

Instituto Superior de Engenharia do Porto

Engenharia Informática



De: Ana Raquel Silva Faria

Orientador: Jorge Morais

2002

Agradecimentos

Agradeço ao meu orientador de projecto, Jorge Morais, pela sua disponibilidade e sugestões durante o decurso do projecto.

Agradeço à minha irmã Ana Paula Silva Faria e ao meu pai Ludovino Gonçalves de Faria sem os quais a realização deste projecto teria sido impossível.

Índice

Agradecimentos	II
Índice.....	III
Ilustrações	V
Tabelas	V
Glossário	VI
1 Introdução.....	12
2 Conceitos Gerais	15
2.1 Redes e Servidores	16
2.2 Natureza do Site.....	19
2.3 O público alvo.....	21
3 Processo de Conhecimento - do Data Mining ao Web Mining	29
3.1 Processo de Conhecimento	30
3.2 Data Mining.....	32
3.3 Text Mining.....	50
3.4 Web Mining.....	53
4 Extracção de Conhecimento na Web	66
4.1 Especificação do Problema.....	67
4.2 Selecção dos dados	69
4.3 Preparação dos dados	72
4.4 Avaliação dos dados	74
4.5 Formatação da Solução.....	74
4.6 Selecção de Ferramentas.....	75
4.7 Construção do Modelo.....	80
4.8 Validação dos resultados.....	81
4.9 Entrega dos Resultados.....	82
4.10 Por os resultados em prática	82
5 Ferramentas	85
5.1 Visão Geral	86
5.2 Blue Martini Marketing.....	87

5.3	<i>Clementine</i>	91
5.4	<i>Megaputer WebAnalyst</i>	96
5.5	<i>MicroStrategy Web Traffic Analysis Module</i>	99
5.6	<i>XML Miner</i>	102
5.7	<i>WebTrends</i>	104
5.8	<i>123LogAnalyser</i>	107
5.9	<i>Webanalyzer</i>	110
Conclusão		112
Referencias		114
Bibliografia		115
Anexos		118
Índice Remissivo		155

Ilustrações

FIGURA 1: INTELIGÊNCIA ARTIFICIAL, SISTEMAS BASEADOS NO CONHECIMENTO	12
FIGURA 2: ETAPAS DO PROCESSO DE CONHECIMENTO	30
FIGURA 3: CRONOGRAMA.....	34
FIGURA 4: BASE DE UM <i>DATA MINING</i>	36
FIGURA 5: CLASSIFICAÇÃO DE UM CONJUNTO DE DADOS.....	37
FIGURA 6: CLUSTERS	38
FIGURA 7: ESTRUTURA DA ÁRVORE DE DECISÃO.....	39
FIGURA 8: ESTRUTURA DE UMA REDE NEURONAL.....	40
FIGURA 9: OLAP.....	42
FIGURA 10: DATAMART.....	43
FIGURA 11: AMBIENTE IDEAL PARA UMA FERRAMENTA DE <i>DATA MINING</i>	48
FIGURA 12: ACESSO À <i>WEB</i>	54
FIGURA 13: ABORDAGENS DO <i>WEB MINING</i>	54
FIGURA 14: ESTRUTURA DE LINKS.....	56
FIGURA 15: FUNCIONAMENTO DAS ASP.....	63
FIGURA 16: EXTRACTO DE DE UM FICHEIRO DE REGISTO DE ACESSOS.....	69
FIGURA 17: DESCRIÇÃO DO <i>SITE</i> [SPILIOPOULOU, 99]	70
FIGURA 18: HIERARQUIA DO CONTEÚDO DAS PÁGINAS [SPILIOPOULOU, 99]	70
FIGURA 19: HIERARQUIA DE UTILIZADORES [SPILIOPOULOU, 99].....	71
FIGURA 20: HIERARQUIA DE PÁGINAS [SPILIOPOULOU, 99].....	71
FIGURA 21: ARQUITECTURA DO <i>WEBANALYST</i>	98
FIGURA 22: ESTATÍSTICAS	104
FIGURA 23: MODO GRÁFICO.....	104
FIGURA 24: MEDIR E MAXIMIZR O <i>ROI</i>	105
FIGURA 25: 123LOGANALYZER.....	108
FIGURA 26: ACCRUE HITLIST	127
FIGURA 27: <i>DATA MINING SUITE</i>	136
FIGURA 28: REGISTO DE ACTIVIDADE	138
FIGURA 29: SUMÁRIO DO PERFIL DO VISITANTE DO <i>NETTRACKER</i>	144
FIGURE 30: CAMPOS RELACIONADOS COM O <i>DATA MINING</i>	149
FIGURA 31: INTERGRAÇÃO DOS DADOS.....	150
FIGURA 32: EXPLORAÇÃO DE REGRAS	151

Tabelas

TABELA I: FERRAMENTAS DE MODELAÇÃO.....	68
TABELA II: TIPOS DE FICHEIROS.....	73
TABELA III :TIPOS DE FERRAMENTAS.....	75
TABELA IV: METODOS DE PROMOÇÃO DO <i>WEB SITE</i>	123

Glossário

A

Algoritmo - É um processo computacional que recebe valores de *input* e produz valores de *output*.

Análise de cesto de compras (*Market Basket Analysis*) – É uma análise de *data mining*, tipicamente feita pelos vendedores, que tratam as compras em número de itens (carrinho de compras) como sendo uma única transação com o intuito de descobrir e explorar padrões de compras.

Aprendizagem automática (*Machine Learning*) - Ramo de Inteligência Artificial que trata com o desenvolvimento de aplicações de algoritmos de aprendizagem automática.

Armazéns de dados – Um sistema de suporte à decisão para armazenar grande quantidade de informação.

Árvores de Decisão – Representação gráfica das relações entre as variáveis dependentes (*output*) e um conjunto de dados de variáveis independentes (*inputs*). As árvores podem ser binárias ou com múltiplo ramos, dependendo do algoritmo utilizado.

ASCII – Acrónimo de *American Standard Code for Information Interchange*. O *American National Standard Institute* estabeleceu um código de caracteres para a transferência de texto entre vários sistemas.

B

Base de Dados – Colecção de dados é guardada numa unidade. As bases de dados são úteis para guardar e tornar disponível uma grande quantidade de dados. Dentro das bases de dados, os dados podem ser organizados em diferentes tabelas, compostas por linhas e colunas.

Branding - Reconhecimento da marca. O utilizador pode não ter seleccionado o *banner*, mas agora ele sabe o nome do seu produto e da sua empresa, e pode ser um futuro cliente.

C

C5.0 - Um algoritmo de aprendizagem automática (*Machine Learning*) tem um conjunto de regras que podem ser aplicadas a conjuntos de dados desconhecidos, para fazer uma nova previsão. Este algoritmo foi desenvolvido

por J. Ross Quinlan, sendo o sucessor dos algoritmos C4.5 e ID3 (*Iterative Dichotomiser*).

CART – *Classification And Regression Trees*. Algoritmo estatístico. Técnica utilizada na segmentação de uma base de dados. O algoritmo cria árvores binárias e segmentos em duas partes. Este algoritmo foi desenvolvido por L. Briemen em 1984.

Classificação – O processo de divisão de um conjunto de dados em grupos mutuamente exclusivos

Click Through – Ao entrar num *banner*, acede-se ao *site* do anunciante ou tem acesso a mais informações sobre o produto ou serviço.

Clicks – O número de cliques que foram feitos no *banner*. Pode-se vender uma quantidade determinada de cliques num determinado *banner*. É como um cartão de telefone, quando acabam o número de créditos o cartão deixa de valer. Por exemplo, podem ser vendidos mil cliques. Ao serem realizados mil cliques neste *banner* termina o contrato que pode ser renovado.

Cota – É uma quantidade de CPM que se compra, como no caso do cartão telefónico.

Comércio Electrónico - Comércio electrónico, ou *e-commerce* (*electronic commerce*), é nome dado ao sistema comercial com a capacidade de realizar transacções envolvem a troca de bens ou serviços entre duas ou mais partes de forma automática, utilizando-se da *Internet*.

CPA - Custo Por Acção. Outra forma de cobrança é a CPA, que cobra por operação bem sucedida, ou seja, ao ser realizada uma compra ou efectuado um serviço com origem no clique do banner. A CPA é calculada dividindo-se o produto da multiplicação do custo por cliques pelas operações bem sucedidas ou acções (*actions*). Resumindo $CPA = CPC * Clicks / Actions$

CPC - Custo por clique. É calculado dividindo-se o custo por mil *banners* (CPM) por Clicks, ou seja $CPC = CPM / Clicks$.

CPM - Custo por mil *pageviews* de um banner.

D

Descoberta de agrupamentos (*Clustering*) – É um processo pelo qual um conjunto de dados é dividido em conjuntos ou *clusters*. Difere da classificação porque não tem um número definido de classes. Podem ser mutuamente exclusivos, concorrentes ou hierárquicos.

F

Formulários (*Forms*) – Técnica pela qual um utilizador pode introduzir dados num página *Web*.

H

Histórico – Registo histórico da relação entre consumidor e vendedor.

Hits – O *Hit* é uma requisição de item como por exemplo páginas ou imagens. Ao entrar em um *site*, é gerando um *hit* para a página aberta e mais um *hit* para quantas figuras existirem nesta página. Se esta página possuir 4 imagens, serão gerados 5 *hits*, um por cada imagem, mais um pela página. Isto repete-se a cada nova página que se entra.

HTML (*Hypertext Mark-up Language*) - Linguagem de codificação utilizada para construir documentos de hipertexto, consultados na *WWW*

M

Mercado de dados (*Datamart*) – Uma base de dados que foi desenhada para ser interrogada e para estruturar relatórios. A base de dados de *Datamart* é separada de uma base de dados de transacções. Os dados derivam dos dados de uma base de dados de transacções e podem incluir dados de outras fontes. Pode conter também dados históricos. O objectivo do *datamart* é melhorar a informação para uma segmento particular de utilizadores ou para um departamento de que uma empresa inteira. Ao contrário de uma base de dados de transacções, tem menos tabelas para relacionar, mais índices disponíveis para aumentar a velocidade de acesso aos dados.

N

Negócio Electrónico (*e-Business*) - Diferente do conceito de comércio electrónico. Comércio Electrónico (na *Internet*) diz respeito a transacções comerciais conduzidas através da *Internet*. Negócio Electrónico é uma expressão usada normalmente para descrever o conjunto de procedimentos e sistemas que têm de ser implementados para que a utilização da *Internet* se torne uma competência nuclear da gestão de uma dada empresa e uma característica intrínseca de todas as suas cadeias de valor. O *e-Business* tem um lema: "*customer satisfaction is everything*": **só importa a satisfação do cliente!**

O

ODBC - *Open DataBase Connectivity* é uma especificação projectada pela Microsoft para permitir às aplicações do Windows acesso a múltiplos dados através de um método simples, sem considerar os diversos formatos dos arquivos de dados, Superando o problema das bases de dados diferentes que têm meios diversos de prover acesso para a informação contida neles.

OLAP (*Online Analytical Processing*) - Aplicação de base de dados que permite ao utilizador ver, navegar, manipular e analisar bases de dados multidimensionais.

P

Pageviews – É página que está a ser carregada. Por exemplo, entrou num site. Abriu a página principal, 1 *page view*. E entrou no *link marketing*, segunda *page view*. Se for para outro *link* ou voltar para uma das duas páginas anteriores será a terceira *page view*.

Portal - “Porta de entrada na *Internet*”. Nome que se atribui a um conjunto de Web sites que assumem duas funções aparentemente contraditórias (mas que fazem sentido no universo ligado da *World Wide Web*). Por um lado, são um repositório tão completo quanto possível de informação (sobretudo notícias e crónicas) e serviços (correio electrónico gratuito; grupos de discussão; chat; etc.) próprios. Por outro lado, proporcionam directórios de *links* e/ou serviços de pesquisa para acesso à *Internet*.

Prospecção de Dados (*Data Mining*) – é uma das fases do processo de descoberta de conhecimento que permite extrair padrões ou perfis de grandes bases de dados, utilizando técnicas de Inteligência Artificial, técnicas de estatística e técnicas de *marketing*.

Proxy - um programa de cache colocado no servidor que fica entre o *browser* do utilizador e o *site*.

S

Sistema de apoio à decisão (*DSS - Decision Support System*) - é um sistema desenvolvido para ajudar as empresas no acesso a informações críticas para os negócios, de forma rápida e segura, agilizando as questões relativas com a gestão e tornando a empresa cada vez mais competitiva. Normalmente possuem interfaces gráficas muito amigáveis, além de permitirem um processo de personalização, no sentido de atender as necessidades específicas de cada empresa.

O DSS deve integrar grande parte das informações das empresas, provenientes de várias fontes, tanto dos sistemas, como de informações externas do mercado, entre outras.

SQL (Structured Query Language) - É um *standard* das linguagens de programação, sendo utilizada para interagir com as bases de dados relacionais. SQL permite obter os dados das tabelas e também, inserir, actualizar e apagar dados.

Stopwords – Palavras a serem ignoradas, pois param o processamento da análise.

R

Rede neuronal com propagação retroactiva (*back-propagation*) – É uma arquitectura de rede que está desenhada para utilizar o *input* e o *output*, é uma camada escondida. Durante a fase de treino a informação é propagada para trás na rede, o que permite a actualização dos pesos das conexões.

Redes neuronais de Kohonen – Pode também ser conhecida por *Self-organizing map*. É uma arquitectura neuronal que utiliza a aprendizagem automática para fazer análise de *clusters*. Foi desenvolvida Teuvo Kohonen.

ROI - Retorno do Investimento. Qual o valor máximo em CPC que se deve pagar para ter um retorno

T

Text Mining - É o processo de extrair conhecimentos de grandes volumes de texto (KDT - *Knowledge Discovery in Texts*), não é conhecida a sua dimensão, e são utilizados para apoiar a tomada de decisão e/ou sumariar textos (documentos, artigos de jornais, etc.). O *Text Mining* permite aos utilizadores explorar e inferir sobre grandes conjuntos de texto, descobrindo relacionamentos escondidos no universo textual.

O *Text Mining*, tal como o *Data Mining*, surgiu da intersecção entre várias áreas, principalmente aprendizagem automática, da estatística e das bases de dados.

U

Unique Users – É o número de utilizadores que visitaram o *site* num determinado período.

Neste caso também, *software* especializado irá fornecer a análise do desempenho dos *banners* para que se possa analisar a sua eficiência.

V

Valor contínuo – Valores contínuos são valores no intervalo de números reais. São o oposto de discretos e categóricos.

Vortal (*Vertical Industry Portals*) - Também chamados de “*vertical trade communities*”, comunidades virtuais verticais, definidas em função de um mercado específico, ou ainda de “*Net marketplaces*” (mercados *online*). Os vortais (ou “portais temáticos”) em português, são os lugares de encontro *online* de indústrias inteiras, onde os trabalhadores de um determinado sector de actividade podem obter facilmente informação específica ao sector, notícias sobre a actualidade do segmento, calendários dos eventos previstos, notas sobre os últimos avanços tecnológicos e estatísticas, entrada em grupos de discussão onde se trata de matéria relevante, do ponto de vista profissional, e aceder a serviços orientados para as suas necessidades específicas.

W

Web Mining – *Data Mining* virado para a *Web*. Descoberta de padrões baseada em dados recolhidos na *Internet*.

X

XML (*eXtensible Markup Language*) - Um termo, relativo a uma especificação de segunda geração do padrão HTML. Em termos práticos: o XML possibilitará, através de tags específicos definidos sectorialmente, uma muito maior eficiência nas pesquisas.

XSL (*eXtensible Style Language*)- Linguagem de formatação do XML

1 Introdução

Na última década assistiu-se a um acentuado desenvolvimento das capacidades informáticas de geração e armazenamento de dados. Avanços científicos na forma de recolha de dados, tais como a utilização de sensores remotos ou satélites espaciais, introdução de códigos de barras nos produtos comerciais e a informatização da maioria das transacções comerciais sob a forma de cartões de crédito, contribuíram significativamente para o aumento dos dados. Paralelamente, os avanços da tecnologia de armazenamento de dados, tornam-se mais rápidas, baratas e com maior capacidade de armazenamento, que dando assim suporte ao aumento crescente de dados e de informação.

A importância do conhecimento para as organizações surge pela intergração de sistemas de informação e de ferramentas computacionais que trabalham directamente sobre o conhecimento aparecendo então os Sistemas Baseados em Conhecimento. Os Sistemas Baseados em Conhecimento como o próprio nome indica, baseiam o seu comportamento inteligente no conhecimento que detêm sobre determinado domínio.

Os Sistemas Baseados em Conhecimento são uma subclasse da Inteligência Artificial em que há uma separação clara entre o conhecimento e o raciocínio, ou seja o controlo do programa não se mistura com a especificação do conhecimento.

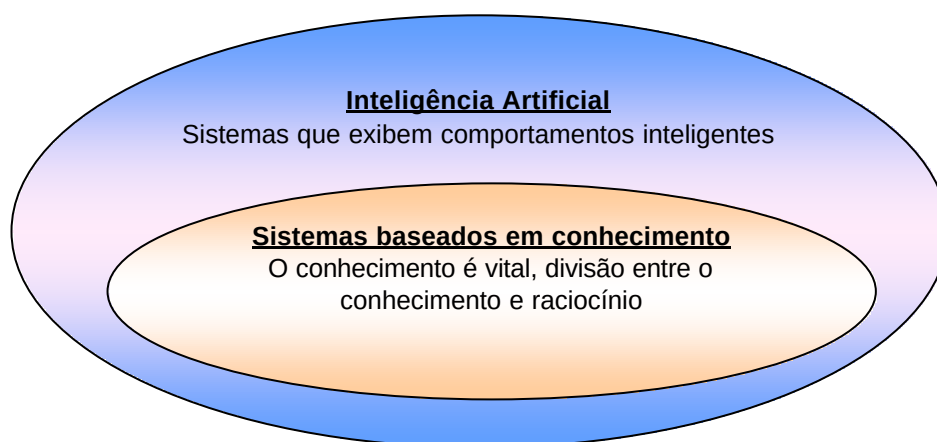


FIGURA 1: INTELIGÊNCIA ARTIFICIAL, SISTEMAS BASEADOS NO CONHECIMENTO

Dos Sistemas baseados em Conhecimento surgiu um variado leque de ferramentas desde as ferramentas de *Data Mining*, aos correctores ortográficos, tradutores automáticos e os agentes inteligentes. Este último tem a sua origem em Sistemas Multi-Agente da Inteligência Artificial distribuída e tem como objectivo povoar a *Internet* auxiliando o utilizador a filtrar *e-mail*, a pesquisar notícias e informação, lembrando marcações, e principalmente como mediadores no comércio electrónico .

Data Mining e a Descoberta de Conhecimento é uma disciplina que envolve o estudo de técnicas que procuram padrões em grandes conjuntos de dados. A

aplicação das técnicas de *Data Mining* à *Web*, é chamado *Web Mining*. O *Web mining* é definido como o estudo das técnicas de *Data mining* que automaticamente extraem informação da *Web*. Estas técnicas permitem aos fornecedores de conteúdos melhorar a qualidade do serviço do *web site*, ou fornecer ao utilizador individual auxílio na navegação e na informação contida no *web site*.

As técnicas do *web mining* podem ser divididas em três campos de pesquisa: *web content mining*, *web structure mining*, e *web usage mining*. *Web content mining* é um campo que foca o desenvolvimento de técnicas que assistem o utilizador a encontrar documentos *web* que estão dentro de determinado critério. O *web structure mining* procura desenvolver técnicas para tirar partido da estrutura de *links* existente nas páginas *web*. Os links são vistos como um mecanismo que dá voto de “popularidade” às páginas que apontam. Finalmente o *web usage mining* foca as técnicas que procuram padrões no comportamento de navegação dos utilizadores da *Web*.



As técnicas de *Web usage mining* permitem compreender as preferências dos utilizadores através do seu padrão de navegação, ajudando a melhorar a topologia de um *site* de acordo como os objectivos de negócio. Esses objectivos podem ir desde da personalização das páginas *web*, a melhorar o tempo gasto no *site*, ou à introdução de páginas novas em lugares que possam ser facilmente vistas.

A população da *Internet* e o crescente progresso da tecnologia *Web* permitem que milhares de documentos sejam todos os dias publicados e acedidos por uma imensa comunidade de utilizadores. Como consequência, os servidores vão acumulando os registos de acessos às suas publicações *on-line*. Existe a percepção do grande potencial destes dados como fonte de conhecimento acerca das reacções e motivações dos utilizadores mas, para os explorar, são necessárias ferramentas de análise adequadas.

Pretende-se neste trabalho verificar qual o actual alcance dos serviços *on-line* baseados em conhecimento partindo da análise de casos conhecidos e da identificação das suas características, verificar quais as linhas de investigação na área e qual o impacto que poderão ter na evolução futura desses serviços.

No capítulo dois são focados os conceitos gerais da disponibilização e acesso a recursos da *web*, a classificação dos mesmos, os sistemas de segurança e todos os mecanismos que permitem atrair e manter os clientes pela *Internet*.

No terceiro capítulo são referidas as principais características, técnicas, algoritmos e aplicações de uma das fases da descoberta de conhecimento o *Data Mining*. São abordados os benefícios e as aplicações do *Text Mining*, processo pelo qual se pode extrair conhecimento de grandes volumes de texto

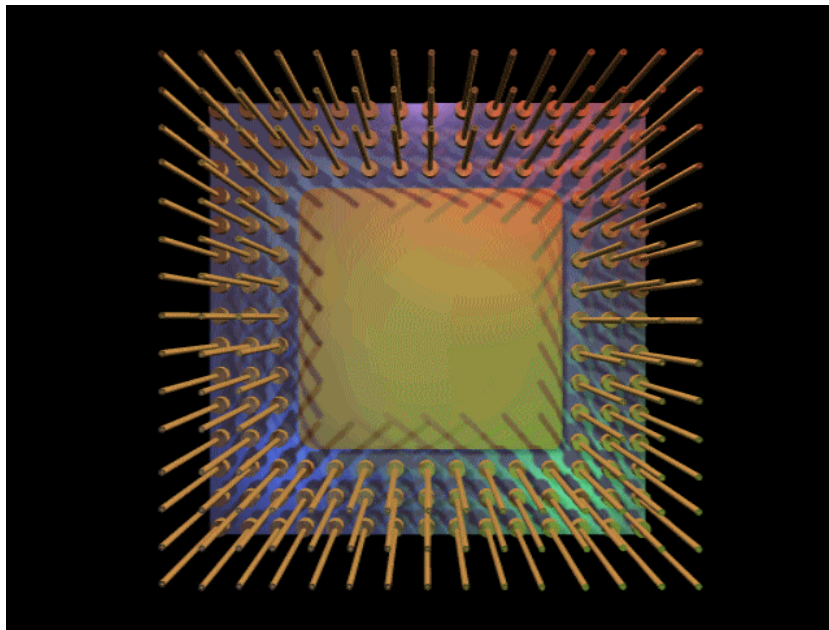
(documentos, artigos de jornal, etc.). E são explicados as abordagens do *Web Mining*: *Web Content Mining*, *Web Structure Mining*, *Web Usage Mining*. Sendo referidas as suas características e áreas de aplicação. Este capítulo também inclui as principais linguagens utilizadas na *Web* e as linguagens para a extracção de conhecimento.

O quarto capítulo descreve todos os passos necessários e as dificuldades em extrair o conhecimento da *Web*, bem como as medidas necessárias para obter resultados fiáveis numa análise.

O quinto capítulo descreve as principais ferramentas existentes no mercado para extrair conhecimento da *Web*. As áreas para que estão vocacionadas desde a análise dos perfis de clientes a análises de associações de produtos.

Sendo assim este trabalho pretende demonstrar a importância das transacções na *Internet* e a importância das técnicas de extracção de conhecimento através de dados retirados da *web* e o seu impacto futuro nos diversos sectores de actividade.

2 Conceitos Gerais



“Os conceitos gerais da disponibilização e acesso a recursos da Web, a classificação dos mesmos, os sistemas de segurança e todos os mecanismos que permitem atrair e manter os clientes pela Internet”

2.1 Redes e Servidores

Uma rede local de computadores – *Local Area Network* (LAN) – é composta por servidores de dados, servidores de programas, servidores *Web*, computadores com a função de estações cliente, equipamentos que permitem a troca de dados entre os computadores da rede com taxas

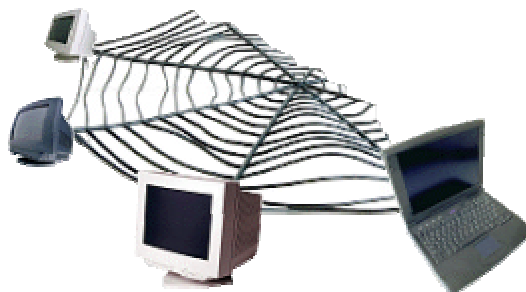


elevadas de transmissão e *software* que permite o a partilha entre computadores e periféricos. Dentro do conceito tradicional de LAN esses equipamentos devem estar restritos a uma área geográfica. Com o avanço da tecnologia essa área está a expandir-se cada vez mais.

Os servidores são computadores robustos, com um ou mais processadores, com grande capacidade de memória RAM e área em disco disponível. Podem trabalhar numa modalidade chamada de *clustering*, que permite que vários computadores possam trabalhar como se fossem um único. No caso de algum deles apresentar uma falha os outros assumem o processamento sem impacto visível.

Os servidores de dados armazenam as bases de dados (Oracle, Informix, SQL Server e outros), nos servidores de programas residem os programas que atendem as estações clientes da LAN e os servidores *Web* disponibilizam *software* e as facilidades da *World Wide Web* (WWW) para as redes *Internet* e *Intranet* (Netscape e Microsoft Web Servers). As estações clientes são os computadores utilizados pelos utilizadores.

Entre os componentes que permitem a transmissão de dados entre os computadores, destacam-se os *routers*, *hubs*, *Ethernet Switches*, cabos e conectores especiais. Os *routers* são equipamentos que direccionam uma informação de uma LAN para outra LAN, através de um protocolo de comunicação específico. Os *hubs* são equipamentos onde são conectados todos os computadores de uma LAN através de cabos e conectores especiais. Os *routers* ligam as LAN também através dos *hubs*. Os *Ethernet Switches* são comutadores de taxas de transmissão e protocolos permitindo a troca de informações em alta velocidade entre dois pontos, normalmente utilizando fibra óptica.



Para apoiar toda essa estrutura de comunicação é necessário de *software* de rede, conhecido como sistema operativos de rede – NOS(*Network Operating System*) – que permitem a partilha de recursos da rede, tais como controle de acesso aos dados armazenados nos servidores, manter de segurança de acesso aos dados

somente para pessoas autorizadas e partilha de impressoras entre outros recursos.

A troca de informações entre os computadores e periféricos está baseada num mecanismo de endereçamento onde cada componente da rede possui um único número, conhecido como endereço IP - *Internet Protocol*.

2.1.1 ISP - Internet Service Provider

As empresas que disponibilizam o acesso aos recursos da Web aos utilizadores finais são denominadas tecnicamente por ISP (*Internet Service Provider*).

Permitir o acesso à *Internet* é simplesmente interligar um computador, ou mesmo uma rede de computadores, à *Internet*, permitindo a comunicação em todos os outros computadores que também estejam ligados à rede.

Esse acesso pode ser de dois tipos:

Utilizadores individuais: Em geral pessoas físicas que se conectam à *Internet* por diversos objectivos, desde utilizar recursos de correio electrónico até divulgação de serviços pessoais. Normalmente o seu acesso é do tipo *dial-in*, entre seu computador pessoal e as instalações de um ISP.



Por utilizadores Institucionais(linhas dedicadas): São empresas que conectam parte ou toda a sua rede à *Internet*, com o objectivo de fornecer acesso à *Internet* para os seus funcionários, utilizar a *Internet* como meio de comunicação entre filiais e clientes, ou mesmo prática de comércio através da *Internet*.

Para que o utilizador possa aceder à *Internet* é necessário que através do seu computador, possa ligar ao ISP. Quando o servidor de comunicação atende é estabelecido um canal de comunicação entre ambas as partes. Essa comunicação é mantida através de uma central telefónica. Nesse processo de conversa, o utilizador identifica-se para o servidor (com *login* e *password*), e ganha um endereço IP para poder navegar na rede.

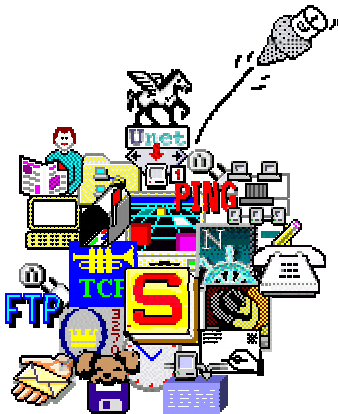
As informações que deseja enviar ou receber através da *Internet*, podendo ser nesse caso, uma mensagem ou uma solicitação de um site, pode ficar na rede local (no servidor) ou sair pelo canal de comunicação do servidor para a *Internet*.

Esses dados que saem de um dos servidores são enviados para o router, que realiza o encaminhamento das informações até outros pontos até que possa alcançar o seu destino.

Antes de chegar ao destino, o pedido pode passar por vários pontos (*router*). A quantidade de pontos depende da localidade do site requisitado.

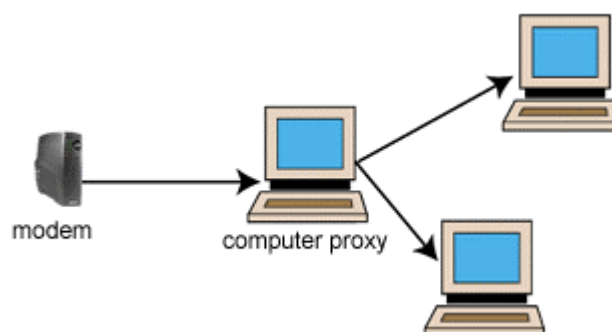
Os serviços extras que o fornecedor de serviços pode prestar são:

E-mail - São serviços que permitem a troca de mensagens entre utilizadores através da *Internet*. São os serviços de maior alcance da *Internet*, pois permitem a troca de mensagens tanto com utilizadores de outras redes de serviços como com utilizadores de redes internas não totalmente interligadas à *Internet*. O funcionamento desses serviços tem como base um endereço conhecido como endereço de correio electrónico (*e-mail address*), cujo formato é apelido@domínio, onde "apelido" representa a identificação de uma caixa postal para recebimento de mensagens e domínio representa o nome do domínio do equipamento que pode localizar essa caixa postal.



Servidor FTP o serviço FTP (*File Transfer Protocol*) é o serviço padrão da *Internet* para a transferência de arquivos entre computadores. A partir dele utilizadores podem obter ou enviar arquivos de ou para outros computadores da *Internet*.

Servidor Proxy servidor específico que arquiva periodicamente as páginas mais acedidas pelos os utilizadores do servidor, acelerando assim a sua navegação e economizando o encaminhamento do *link Internet*.



WebMail Permite a visualização de e-mails pela *Web*, em qualquer parte do mundo, sem necessidade de configuração ou instalação de novos programas no computador em utilização.

2.2 Natureza do Site

O site é a parte estratégica principal para fazer negócios *on-line*. Determinar a natureza do site é definir se o negócio que se quer implantar na Internet será uma loja, um site que fornece informações, *business to business*, *business to consumer*, etc.

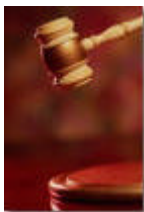
Quanto à natureza os sites podem ser classificados como:



Institucionais Sua finalidade é apresentar a empresa ao visitante. Não realizam vendas e as informações que fornecem, em geral, são apenas sobre a empresa. É muito utilizado por empresas que apenas querem marcar presença na *Web*.



Lojas Virtuais São destinado à vender produtos ou serviços. Podendo ser do tipo *business to business* ou B2B (empresas que vendem produtos ou serviços para outras empresas) ou *business to consumer* ou B2C (vendem produtos directamente ao consumidor).



Leilão *on-line* Tem por objectivo intermediar a compra e a venda de mercadorias na *Internet*, por meio de leilão. Os produtos oferecidos são relacionados por categoria, em muitos casos com a fotografia do produto, que pode ser desde uma moeda antiga até um automóvel. Os compradores escolhem o produto e fazem sua oferta. Num período estipulado de tempo, a melhor oferta leva o produto. Alguns leilões cobram uma taxa do vendedor (em média 5% do valor do produto) outros não cobram taxa.



Cooperativas ou grupos de compradores São sites criados para unir os utilizadores da *Web* que tenham um interesse comum de adquirir um determinado bem ou serviço. O utilizador liga-se ao grupo de seu interesse e quando esse atinge um número significativo de utilizadores, a compra do bem é realizada. A vantagem é a redução no preço do produto, pois geralmente essa compra é feita por atacado, muitas vezes directamente na indústria ou com o produtor. Verifica-se actualmente o aparecimento de grupos de compra formado por empresas.



Sites de Informação Estes tem por objectivo fornecer informações, que geralmente são gratuitas. A facturação destes sites costuma estar no espaço vendido para a publicidade. Empresas compram espaço, geralmente na forma de *banners*. O que determina o valor da publicidade num site de informação é o volume de tráfego (quantidade de *page views*). Dentro desta categoria sites de informação existem especializações, por exemplo: sites de pesquisa (*Yahoo*, etc.), sites de notícias, sites de *download* (*tucows*, etc.), sites de cursos *on-line*, sites de recrutamento de pessoal, etc.

O processo para se estabelecer uma empresa na *Internet* que comercialize bens e serviços é similar ao de uma empresa convencional.

Seja qual for a natureza do *site*, estes devem estar sempre actualizados, principalmente as Lojas Virtuais, que estão a competir com a concorrência. Conhecer bem o cliente e ter uma base de dados eficiente para auxiliar no processo de apoio à decisão são factores muito importantes para se obter vantagens competitivas.

2.2.1 B2B - Business to Business

Business-to-business é o nome que se dá a transacções comerciais entre empresas. Muitas empresas utilizam este tipo de *e-commerce* para se relacionar com seus fornecedores, fazendo pedidos, recebendo e pagando facturas e pagamentos, trocando dados, captando novos parceiros, etc. Esta relação entre duas ou mais empresas, envolvem tecnologias que utilizam o Intercâmbio de Dados Electrónicos (IDE). O IDE permite a comunicação directa entre os sistemas informatizados das empresas sem precisar da interferência do homem.

2.2.2 B2C - Business to Consumer

Outra forma de comércio na *Internet* é o *business-to-consumer*. Nesta categoria, a empresa vende directamente ao consumidor.

Um determinado consumidor entra no seu *site* em busca de um produto. Pesquisa por informações referentes a este produto, manual técnico, funcionamento, preço, prazo de entrega, assistência técnica, etc.

Satisfeito com as informações, ele começa a realizar a compra. Entra no módulo adequado para realizar esta compra, escolhe o produto ou produtos.

O próximo passo será escolher como vai pagar este produto. Ele faz uma opção, por um determinado cartão de crédito. Carrega no ícone deste cartão, preenche os dados e envia. No prazo estipulado ele recebe no local indicado o produto da compra.

Este tipo de comércio permite manter baixos *stocks* de produtos, que podem ser montados ou adquiridos na medida em que são vendidos.

2.3 O público alvo



Conhecendo-se a natureza do site, determina-se o público alvo. Este pode ser definido por segmentos (indústria, comércio, serviços, instituições públicas, etc.), por grandes áreas (educação, medicina, engenharia, economia), de todas as idades (público jovem, terceira idade, adulto), preferências (coleccionadores, aficcionados do desporto, amantes da culinária) e assim por diante. Dependendo da natureza e do público alvo, direcciona-se o conteúdo do site.

2.3.1 O conteúdo do Site

Nessa fase deverão ser definidos quais produtos ou informação que vão estar disponíveis do site.

No caso de uma loja virtual, sites de leilão ou grupos de compra é o momento de se caracterizar os produtos que serão oferecidos e determinar e informações sobre estes estarão disponíveis. É conveniente colocar dados técnicos, preço, garantia, prazo de entrega, assistência técnica, telefone para contacto (caso o cliente queira alguma informação adicional).

No caso de vendas ou leilão, as fotografias dos produtos são muito importantes. Mas ter em conta, que imagens gráficas ou fotografias demoram muito tempo a carregar, que torna o site lento abrir. Pesquisas revelam que o tempo médio de espera para abertura de uma página, para um cliente é de 5 segundos.

Se o site for de natureza informativa, é hora de definir as informações que estarão disponíveis. Lembrando que, em ambos os casos, é imprescindível colocar o *e-mail* para contacto.

2.3.2 A Informação

O que atrai as pessoas para a *Internet*? A resposta é: Informação. A Informação é a moeda da *Internet*. Não existe nada mais rico em informação do que a *Web*.

A informação presente na *Web* pode ser encontrada de diversas formas. Se quer ir ao teatro, basta entrar em alguns *sites* para saber tudo sobre as peças em cartaz, desde os actores até o preço dos bilhetes, podendo inclusive compra-los *on-line*. Se um estudante necessita de pesquisar sobre um tema, lá está num website. Informações sobre história, turismo, religião, indústria e comércio, cursos dos mais diversos (inclusive *on-line*), geografia, música etc.

Existe ainda um tipo de informação bastante valiosa que é a informação de perfil. Quantas vezes, ao fechar um *site* ele dispõe a oferecer algum serviço

(em geral gratuito). Então, responde-se a um questionário, que pergunta as nossas preferências, habilitações, salário, etc. Essa informação é vital para a sobrevivência das empresas. Qual o empresário que não gostaria de saber o perfil do seu cliente: o que compra, como compra e quando compra? Quais suas preferências, como encontra-lo. Mesmo sem preencher questionários é possível saber outros dados como o horário de mais acesso à *Internet*, os sites de maior preferência, etc.

O preço dos produtos e serviços

O preço dos produtos na *Internet* devem, de uma forma geral, ser mais baratos do que no mercado convencional, uma vez que não existe necessidade de se manter grandes *stocks* ou um grande número de funcionários.



Ao estabelecer os preços, é aconselhável que o seu preço seja mais atractivo do que o do seu concorrente, agora está “a um *clique* de distância”, pois os clientes certamente irão procurar as melhores ofertas.

O pagamento

As formas de pagamento mais utilizadas são a transferência bancária, cartão de crédito, depósito bancário, pagamento nos correios e outros. Um factor que tem mais importância é a segurança nas transacções. Devido aos frequentes ataques de *hackers*, o consumidor não se sente seguro em colocar os seus dados na rede, nas compras *on-line*. Para a segurança da página existem alguns protocolos como SET (*Secure Electronic Transaction*), SSL (*Secure Socket Layer*), criptografia, utilizados para realizar transacções seguras na *Internet*.

Segurança

Uma das grandes barreiras para a realização de compras na rede é a questão da segurança na transmissão dos dados, principalmente dados referentes ao cartão de crédito.



Esta preocupação por parte do utilizador tem fundamento, uma vez que as transmissões de dados através de redes de comunicação públicas, como é o caso da *Internet*, muitas vezes são vítimas de falhas técnicas e humanas ou interceptados por *hackers*, o que gera grandes problemas aos *sites* e aos utilizadores de serviços virtuais.

Tudo isso pode ser evitado através da utilização de algumas tecnologias que fornecem excelentes níveis de segurança. De uma forma geral, o problema não está na transmissão de dados propriamente dita. Ao colocar o número do seu cartão de crédito num formulário de compras num *site* e envia-lo ao local requerido, muitas vezes o utilizador imagina que esta transmissão

possa ser interceptada por alguém durante o percurso dos dados pela *Internet*.

Isto não acontece. As falhas de segurança ocorrem no local de armazenamento do *site*, ou seja, o utilizador digita o número do seu cartão e envia para a empresa fornecedora do produto. Esta empresa recebe o número deste cartão na base de dados. É aí que está a falha de segurança. Portanto, a empresa que recebe dos dados deve possuir um sistema de segurança adequado e eficiente.

Sistemas de segurança

O protocolo SET

SET (*Secure Electronic Transaction*) é um protocolo seguro, desenvolvido em 1997 pela empresa de cartão de crédito Visa, em parceria com empresas de *software* e de *Internet*, com o objectivo de implementar segurança nas transacções electrónicas envolvendo o pagamento com os cartões de crédito.

O SET está baseado na distribuição de assinaturas digitais para as partes envolvidas na transacção, evitando-se assim que seja divulgado o número do cartão de crédito. As assinaturas digitais são usadas para autenticar todas as partes envolvidas na transacção.

O site e o consumidor recebem as assinaturas. Isto faz com que ambas as partes estejam certas de que estão a comunicar comercialmente com partes reais e idóneas. Do lado dos *sites* existe o Servidor do comércio, que é onde ficam armazenados as assinaturas digitais das lojas sob a responsabilidade da empresa de cartão de crédito.

Criptografia

Protocolo SSL

O protocolo SSL (*Secure Socket Layer*) foi criado pela Netscape para tentar solucionar o problema de segurança que envolve as transacções com cartão de crédito.

As informações transmitidas são codificadas (encriptadas) para que somente o utilizador e o servidor possam interpretar seu conteúdo, assegurando a privacidade da transacção.

Ao visitar um *site*, o sistema envia ao Cliente uma assinatura digital atestando a autenticidade deste site, ou seja, que ele é seguro e que o cliente pode enviar informações que elas serão lidas apenas pelo receptor. O mesmo acontece no caso inverso, assegurando ao site que este visitante existe.

Assinatura Digital

A assinatura digital, ou Certificado de Identidade Digital, pode ser considerado a identidade virtual. Esta identidade virtual pode ser utilizada toda a vez que for necessário identificar-se com absoluta segurança clientes ou empresas. De um lado certifica a existência e autenticidade do cliente e do outro certifica que este cliente estará fazer negócio com sua empresa e não com uma empresa falsa.

Este certificado é emitido por uma CA (Autoridade Certificadora Digital – *Certificate Authority*)

A CA atribui identidades a duas chaves electrónicas contendo nome do proprietário, data de expiração deste certificado, assinatura, número série, etc, que podem ser utilizadas para encriptar ou assinar informações digitais.

2.3.3 Formulários

Os formulários são espaços em branco para preencher que existem nos *websites*. Eles possuem diversas funções:

Servem para recolher dados e informações que serão preenchidos pelos utilizadores. Alguns exemplos dessas informações podem ser:

- Pedido de compra de produtos;
- Pedido de informações adicionais sobre um produto ou serviço,
- E-mail por permissão (o utilizador solicita a recepção de informação por e-mail),
- Pesquisas on-line sobre um tema específico (geralmente uma pergunta que fica no site por um período determinado - por exemplo: Tem cartão de crédito internacional?),
- Perfil do utilizador que visita o site (em geral, um longo questionário - nessa modalidade sempre há uma troca, o utilizador coloca as suas informações e a empresa fornece grátis algum serviço ou produto de seu interesse).

Após o preenchimento, o formulário pode ser enviado para uma conta de e-mail da empresa (opção mais simples) ou ir directamente para uma base de dados, que poderá estar integrada num sistema mais sofisticado de gestão, que analisará esses dados e os encaminhará para efectivar a transação.

Existem diversas tecnologias próprias para a implementação de formulários on-line, dentre elas existem as linguagens de scripts CGI (*Common Gateway Interface*), muito utilizados em formulários via e-mail. Os códigos ASP (*Active Server Pages*) que podem interligar formulários à bases de dados.

2.3.4 O Marketing

Como em todos os negócios, o comércio electrónico necessita de uma boa divulgação. No caso dos *Web Sites*, o primeiro passo é inscrever a *Home Page* em sites de pesquisa.

Sites de pesquisa são sites que desempenham o mesmo papel de uma lista telefónica, com a diferença de que em alternativa de folhear, o utilizador irá escrever em um formulário uma palavra chave onde a pesquisa se irá basear, o utilizador também poderá procurar por categorias.



A colocação de *banners* é outra forma de divulgação. *Banners* são publicidade paga colocada em outros *sites* que ao serem seleccionadas dão acesso ao site da empresa dona deste *banner*. Mas em matéria de divulgação até agora nada supera os *media* tradicional. Jornais, revistas, rádio, televisão continuam a ser importantes para captar novos clientes e fixação da marca. Em geral, cerca de 60% dos investimentos num *Web site* são destinados à publicidade e *marketing*.

Os negócios virtuais criaram aquilo que se chama de “Nova Economia”. Esta nova forma de relacionamento comercial vem propor uma nova forma de empresa, de cliente, de visão do negócio e, portanto, de *marketing*.

Marketing quer dizer “mercado em movimento”, e nada pode ser mais movimentado do que o mercado virtual. O que se tem como verdade absoluta, pode amanhã ser classificado como obsoleto.

O comércio electrónico está ainda a dar os seus primeiros passos. Todas as tentativas de se movimentar nesta economia podem ser consideradas experiências. Porém, apesar de sua pouca idade, já começa a coleccionar histórias de sucesso e fracasso. E é com base nestas histórias que surgirão novas tentativas.

Por ser novo, uma das grandes dificuldades do *marketing* digital é a de conseguir números que atestem a eficiência deste ou daquele sistema, simplesmente porque ainda não se conseguiram sistemas de recolha de informações totalmente eficientes. Entre erros e acertos tenta-se criar um tipo de marketing eficiente para o comércio digital.

Acesso a especialistas:

- pesquisa por peritos nas suas áreas de interesse;
- contacto e contratação destes peritos;
- obtenção de informações pagas ou gratuitas;

Avaliação de novos mercados:

- ↻ a *Internet* como fonte de *feedback* imediata;
- ↻ influência da *Internet* na cultura moderna;
- ↻ a *Internet* funciona como uma extensão dos "Media convencional";

Vendas via *Internet*

- ↻ Vender é permitido na *Internet*;
- ↻ Um método de expandir a presença no mercado;
- ↻ É possível modificar o resultado das vendas *on-line*;
- ↻ O que é vendido com sucesso na *Internet* hoje: - *Software (shareware)*, livros e revistas, música (CD, discos);
- ↻ O que pode ser vendido com sucesso - serviços de impressão remota, produtos de consumo, serviços legais, serviços financeiros, venda de imóveis, turismo, etc;

2.3.5 Por que anunciar na *Internet* ?

Apesar de ainda ser pequeno em relação ao seu potencial total, o número de utilizadores no mundo cresce de forma assustadora. A *Internet* já está presente em sistemas de telefone móvel (*wireless*), onde pode-se aceder aos bancos e contas de *e-mail* e, em breve fará parte do dia-a-dia doméstico, com acesso inclusive de monitores instalados nas portas dos frigoríficos.

Esta forma de publicidade possui preços atractivos, se comparadas a outras formas como a televisão, facilidade de produção se comparada ao *marketing* convencional que torna possível o anúncio de empresas de diversos tamanhos entre outras.

2.3.6 Banner

Banner é uma forma de publicidade, em geral de forma rectangular, que fica numa página *Web*. Esta propaganda pode ou não estar associadas a um *link* que abrirá a página do produto ou serviço que está sendo anunciado ou conterá mais informações sobre o produto. Actualmente, o *banner* é a forma de publicidade mais utilizada na *Internet*.

O interesse dos anunciantes vai depender dos objectivos deste. Alguns produtos necessitam de exposição em massa, outros de exposição segmentada. Algumas vezes esta exposição segmentada torna-se mais cara, apesar de ser vista por um número maior de pessoas. Por exemplo, o fornecedor de equipamento médico. Não seria interessante para ele colocaro seu *banner* num grande portal por exemplo. Mesmo este portal tendo milhões de visitas por dia. Seu retorno seria maior se ele pusesse seu produto directo num *site* médico, pois lá está concentrado seu consumidor final. Provavelmente este *site* médico cobrará um valor maior pela exposição do *banner* do que um grande *site* de pesquisa por exemplo.

Mas não pode passar o dia a carregar no botão do rato para que o seu site seja bastante visitado. Isto não funciona. Existem programas que analisam o desempenho dos *sites* fornecendo relatórios contendo números de *Hits* (acessos), *Pager View*, *Click Through*, *user section*, *unique users*, etc.

2.3.7 Como aumentar as visitas ao site

Os seguinte factores podem aumentar a frequência das visitas dos utilizadores a um *site*:

A inscrição nos mecanismos de pesquisa

- Atenção especial na selecção de palavras-chave. Elas devem reflectir a natureza do *site*.
- Actualizar sempre as informações sobre o *site* nos mecanismos de pesquisa.

Resposta imediata a e-mails

- Resposta em menos de 24 horas após a recepção. Não deixa um cliente à espera.
- As respostas devem ser objectivas com um leve toque de *marketing*.

Actualizações de Conteúdo

- Existe sempre o desafio de fazer o cliente voltar ao site. Para isso é necessário que este esteja constantemente acrescentando conteúdo que interesse ao utilizador.

E-mail Directo

- Com um formulário simples e apenas um *e-mail*, é possível criar uma base de dados de *e-mails* para enviar aos clientes que solicitarem as novidades ou actualizações do site.
- Outra coisa que funciona bem é a pergunta "Gostou deste site? Se sim, envie essa página a um amigo".

Promoções e concursos

- Uma pesquisa indicou uma das palavras atractivas aos utilizadores é a palavra "**grátis**". Realizar promoções, promover concursos com prémios, disponibilizar *downloads*, são atractivos que sensibilizam o utilizador a sempre voltar ao *site*.

Divulgação em media convencional

- Anúncios em jornais, revistas, rádio, TVs, brindes e *t-shirts* promocionais, ainda são a maneira mais eficaz de se atingir o público alvo.

Venda de espaço para publicidade

Depois de intensificar o tráfego no seu *site*, o espaço disponível nele será valorizado para a colocação de publicidade. É importante estar atento a possibilidade de vender espaço para publicidade (*banners*).

2.3.8 Onde colocar o site?

A colocação do *site* pode ser vital para uma empresa pelas mais diversas razões desde razões de segurança ao próprio preço das infra-estruturas. Sendo assim existem as seguintes alternativas:

Implantar infra-estrutura

Dependendo da necessidade da empresa essa é uma alternativa viável. O custo para se implantar essa estrutura é alto, envolve a instalações físicas (salas apropriadas), aquisição de equipamentos, a manutenção de técnicos especializados, aquisição de *software*, custos da conexão com a *Internet*. A vantagem é o controle total.

Servidor exclusivo ou inserir um servidor de empresa fornecedora desses serviços.

Em ambos os casos, os custos de manutenção são altos. Contratando um servidor exclusivo a empresa fica à mercê do *software* usados pelo servidor. No outro caso pode utilizar *software* próprio, porém deve arcar com os custos destes. A principal vantagem nesses dois casos é não ter que arcar com a implementação de infra-estrutura na empresa e dos custos com equipas especializadas.

Contratar espaço de colocação num servidor - Web Hosting

A maior parte dos *sites* existentes na *Internet*, são colocados dessa maneira. Nesses casos a infra-estrutura é compartilhada com outros sites no servidor. O custo desse tipo de colocação é baixo. A solução é interessante para pequenas e médias empresas que se iniciam nos negócios *on-line*. Deve-se tomar cuidado ao seleccionar o fornecedor desses serviços. Quanto maior o número de linhas e quanto maior a taxa de transmissão, menor é a ocorrência de congestionamento. Outro cuidado é verificar a existência de uma *firewall* para garantir a segurança do seu *site*.

3 Processo de Conhecimento - do Data Mining ao Web Mining



“As principais características, técnicas, algoritmos e aplicações de uma das fases da descoberta de conhecimento o Data Mining. São abordados os benefícios e as aplicações do Text Mining, processo pelo qual se pode extrair conhecimento de grandes volumes de texto (documentos, artigos de jornal, etc.). E são explicados as abordagens do Web Mining: Web Content Mining, Web Structure Mining. Sendo referidas as suas características e áreas de aplicação. Este capítulo também inclui as principais Linguagens utilizadas na Web e as linguagens para a extração de conhecimento.”

3.1 Processo de Conhecimento

Com o avanço da tecnologia informática e com a descida do custo dos meios disponíveis para a recolha e armazenamento de dados a consequência imediata é um aumento da informação armazenada.

Os dados recolhidos durante um período de tempo contêm conhecimento sobre os negócios ou padrões que caracterizam o comportamento do utilizador.

Segundo [Fayyad et al., 199?] o processo de conhecimento é um processo interactivo e iterativo que envolve vários passos e com decisões a serem tomadas pelo utilizador.

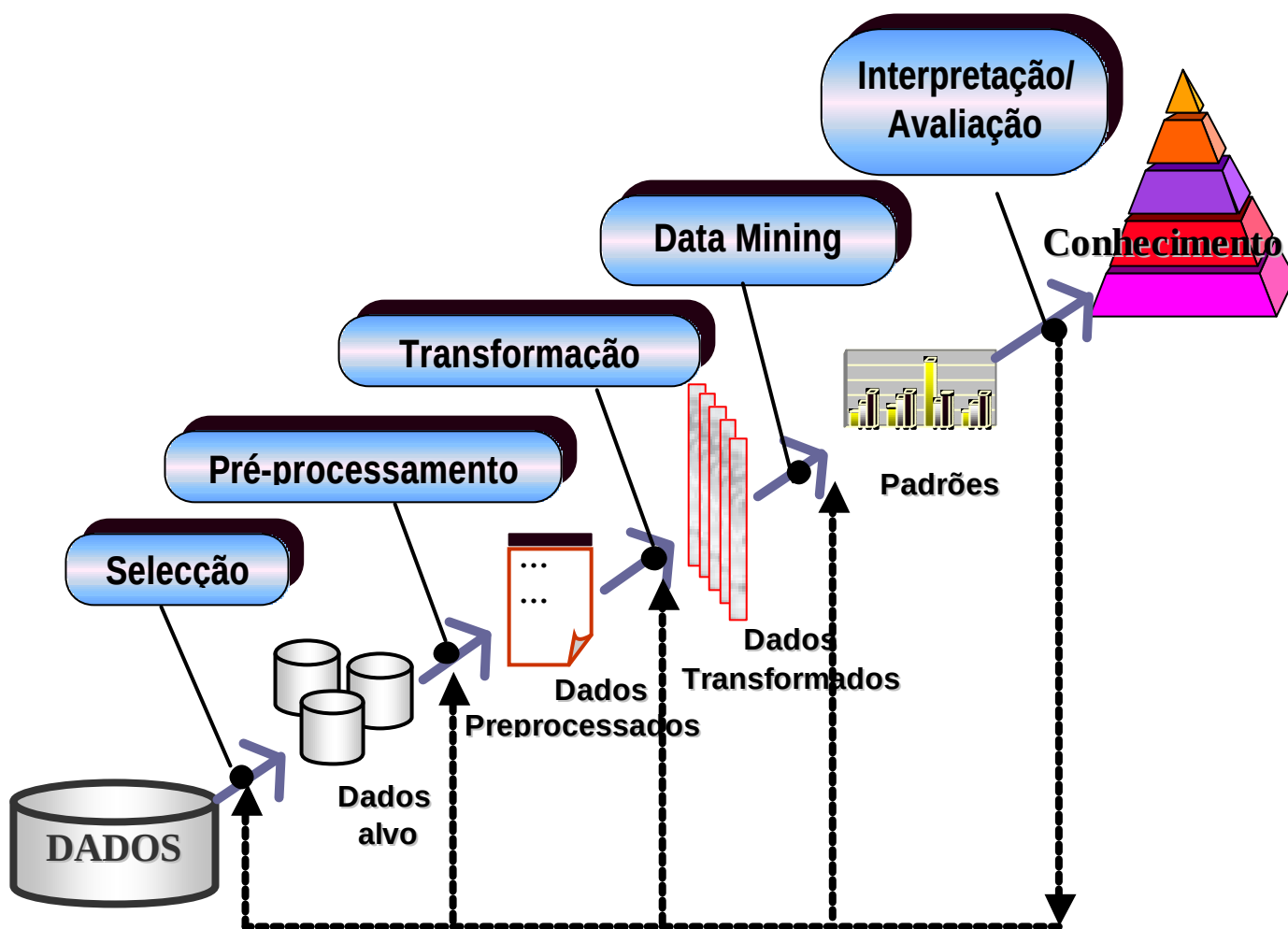
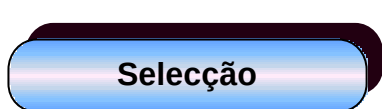


FIGURA 2: ETAPAS DO PROCESSO DE CONHECIMENTO

O ponto de partida para a análise do conhecimento é a percepção do domínio a analisar, conhecimento existente à partida e o objectivo dos utilizadores.



Objectivo: Conjunto de dados alvo

Seleccionar um conjunto de dados, ou um subconjunto de variáveis ou exemplos de dados, sobre o qual se realizará o processo de conhecimento.

Pré-processamento**Objectivo: “Limpeza” de Dados e Pré-processamento**

Operações básicas, por exemplo a eliminação de erros, estratégias para os campos sem informação, controlo das alterações, etc.

Transformação**Objectivo: Redução e projecção dados**

Representação da dependência dos dados de acordo com o objectivo do processo. Utilização de métodos de redução de dimensões ou transformação com vista a reduzir o número de variáveis a considerar e identificar representações sem relevância

Data Mining**Objectivo: Definir finalidade da utilização do método de Data Mining**

Decidir se o objectivo do processo de extração de conhecimento é classificação, regressão, clustering, etc.

Objectivo: Seleccionar o algoritmo de Data Mining

Seleccionar método(s) para pesquisa de padrões nos dados. Decidir os modelos e parâmetros apropriados de acordo com o objectivo do processo de descoberta de conhecimento.

Objectivo: Data Mining

Pesquisar padrões de interesse sob uma ou várias formas de representação: classificação em regras ou árvores decisão, regressão, *clustering*, etc

Interpretação/Avaliação**Objectivo: Análise**

Interpretar padrões obtidos. Possibilidade de retornar a qualquer etapa anterior para optimização dos resultados.

Objectivo: Consolidação do conhecimento adquirido

Incorporação do conhecimento no sistema, ou simples documentação e divulgação para interessados. Deve-se ter a preocupação de validar e resolver conflitos com conhecimento anteriormente obtido.

3.2 Data Mining

Data Mining é uma das fase da descoberta de conhecimento que procura padrões escondidos e em grandes conjuntos de dados. A descoberta de conhecimento em bases de dados é definido como um processo de identificação de padrões válidos, potencialmente úteis e compreensíveis, nos dados analisados.

As tradicionais análises de dados são úteis para criar relatórios a partir dos dados que confirmam as hipóteses predefinidas sobre eles. No entanto, os grandes volumes de dados que estão a ser recolhidos criam novos desafios a essas técnicas, com o mundo de negócios à procura de novas formas de utilizar esses dados como uma vantagem sobre a concorrência.

Ao longo dos anos as organizações acumularam milhares de dados contendo informação, que ajudaram a evoluir e conquistar mercado, permitindo que os administradores, baseados nelas, de diferentes formas, tomassem decisões.

Data Mining veio melhorar a "perspicácia" e a "visão do negócio" dos gestores, possibilitando conquistar e manter clientes com maior facilidade, procurando encontrar respostas para assuntos/situações futuras a partir das situações passadas.

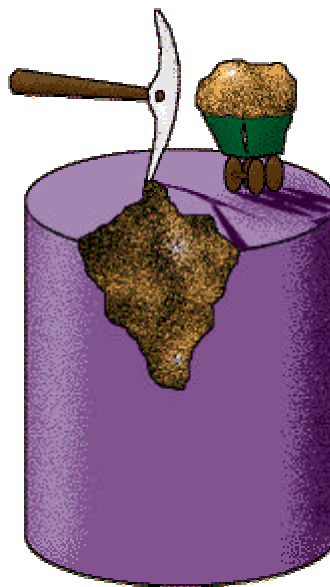
Data Mining prevê acções dos consumidores, mostrando relações entre dados, antes escondidos, descobre hábitos de consumo, melhora os lucros da empresa e a ajuda esta a consolidar-se no mercado.

Empresas que detêm e/ou fornecem com fiabilidade, rapidez e de forma organizada este produto, têm grandes hipóteses de permanecerem de forma competitiva no mercado.

Acerca destas informações, os gestores desejam saber: O que querem os nossos clientes? Como anda a concorrência? Que assuntos causam mais impacto na sociedade? Qual o direcção do orçamento?

Estas inquietações são necessárias para manter a empresa no mercado, para organizar as sua informação.

Uma das soluções encontradas para colocar ordem nos dados foi a criação de armazéns de dados, soluções capazes de armazenar enormes quantidades de dados, integrando-os em áreas lógicas, de assuntos de negócios, armazenando-os de forma a que fiquem acessíveis e compreensíveis para as pessoas responsáveis pela tomada de decisões da empresa.



Surgem daí os instrumentos tradicionais de suporte à decisão que estão preparados para fornecer e discriminar informação, de acordo com as perspectivas escolhidas pelo gestor do negócio, de forma simplificada, em visão tridimensional, centralizando, e organizando enormes quantidades de informação histórica do negócio da empresa.

Todavia não basta ter informação em quantidade, qualidade e organizada. O gestor necessita hoje de respostas para perguntas, tais como: Como obter o máximo de informação útil para minha empresa? Como descobrir padrões de dados e novos conhecimentos? Como manter o meu cliente? Como utilizar adequadamente e descobrir ligações entre eventos nas minhas bases de dados?

Para encontrar respostas para questões como estas, as empresas têm usado as ferramentas de *Data Mining*. Estas Ferramentas permitem prever tendências e comportamentos futuros, permitindo aos gestores tomarem decisões baseadas em previsões consubstanciadas e não em suposições. A análise automatizada e antecipada oferecida por *Data Mining*, vai muito além da simples análise de eventos passados, o que já é fornecido pelas ferramentas de retrospectiva típicas de sistemas de apoio à decisão (como SQL - *Structured Query Language* e OLAP- *On-Line Analytical Processing*).

Por esta característica, das técnicas de *Data Mining* são extremamente adequadas para analisar estes grupos de dados, que seriam difíceis de serem analisados utilizando outras técnicas.



3.2.1 Principais Características de uma ferramenta Data Mining

Os paradigmas de negócio fizeram evoluir e mudar as questões a serem respondidas pelos Sistemas de Informações, ao longo do tempo. Por exemplo:

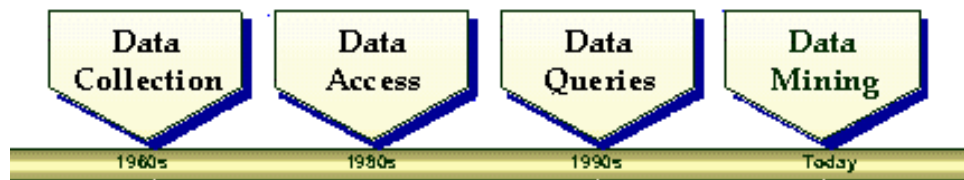


FIGURA 3: CRONOGRAMA

- *Data Collection* (1960s) - "Qual foi o total de vendas nos últimos cinco anos?"
- *Data Access* (1980s) - "Quanto rendeu a nossa filial B no mês passado?"
- *Data Warehousing & Decision Support* (1990s) - "Quanto rendeu a nossa filial B no mês passado? O que isto afectou na filial C?"
- *Data Mining* (Hoje) - "O que realmente irá acontecer com as vendas na filial C no próximo mês? Por quê?" [THEARLING, 2000]

Data Mining é uma técnica que permite pesquisar em grandes bases de dados, informação, que aparentemente, esta camuflada ou escondida, permitindo, com isso, agilidade na tomada de decisão.

Uma empresa que utiliza as técnicas de *Data Mining* é capaz de :

- Criar parâmetros para entender o comportamento do consumidor;
- Identificar afinidades entre as escolhas de produtos e serviços;
- Prever hábitos de compras;
- Analisar comportamentos habituais para se detectarem fraudes.

As ferramentas de *Data Mining* podem determinar padrões de comportamento, como associação de produtos durante um processo de compras. Com isto, as ferramentas *Data Mining* indicam a presença de oportunidades e de previsões em relação àquele público consumidor. O diferencial *Data Mining* está no facto de as descobertas de padrões de consumo se darem por uma lógica de algoritmos com base numa rede neuronal de raciocínios. São ferramentas de descoberta matemáticas, feitas sobre os registos da organização já processados, contra descobertas empíricas.

Por exemplo:

- A rede americana *Wall-Mart*, pioneira no uso de *Data Mining*, descobriu ao explorar seus números que 60% das mães que compram uma boneca Barbie, levam também uma barra de chocolate;
- O banco *Itaú* conseguiu aumentar sua taxa de retorno do envio directo para 30% reduzindo a conta do correio a um quinto, armazenando e analisando a movimentação financeira de seus 3 milhões de clientes nos últimos 18 meses.
- A *Sprint*, um dos líderes no mercado de americano de telefone de longa distância, desenvolveu, com a análise de *Data Mining* no seu armazém de dados, um método capaz de prever com 61% de segurança se um consumidor trocava de companhia telefónica dentro de um período de dois meses. Com um marketing agressivo, conseguiu evitar a deserção de 120000 clientes e uma perda de 35 milhões de dólares em facturação.
- O governo de *Massachusetts*, no Estados Unidos, compilava informações financeiras imprimindo telas e mais telas nos terminais de grande porte. Só com papel foram economizados 250000 dólares por ano, além de permitir que o orçamento anual fosse assinado e analisado, antes do início do ano fiscal.
- O SERPRO no Brasil, implantantou um armazém de dados e *Data Mining*, já consegue hoje cruzar e analisar informação em cinco minutos, o que antes obrigavam quinze dias de trabalho.
- E o clássico exemplo de uma grande rede de vendas americana que descobriu, através das técnicas de *Data Mining*, que as vendas de fraldas estavam intimamente ligadas às vendas de cerveja. Explicação, os pais que saiam à noite para comprar fraldas, compravam cerveja também.

Por isso, e por outros inúmeros exemplos das técnicas de *Data Mining*, este conceito cresce em importância, principalmente nas empresas preocupadas com o futuro do negócio. As técnicas de *Data Mining* acompanham o evolutivo mundo dos armazéns de dados.

3.2.2 Base das técnicas de Data Mining

As técnicas de *Data Mining* estão baseadas na Estatística, em Inteligência Artificial e em aprendizagem automática, conforme apresentado na figura 1:

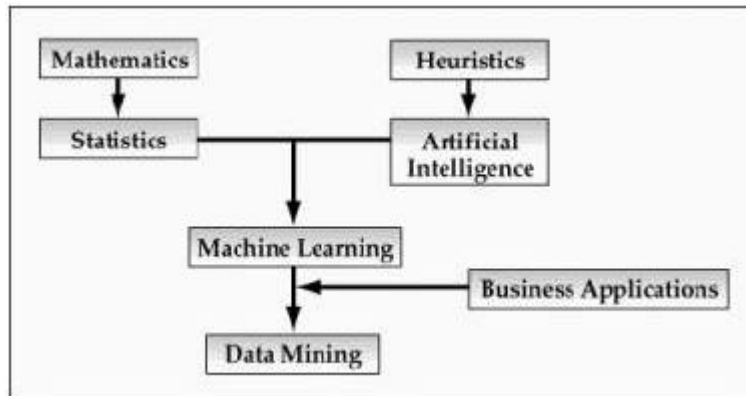


FIGURA 4: BASE DE UM *DATA MINING*

A estatística

É a base de todas as outras tecnologias criadas para *Data Mining*. Conceitos como distribuição normal, variância, análise de regressão, desvio simples, análise de conjuntos, análises de discriminantes e intervalos de confiança, são utilizados para realizar as pesquisas nos dados, bem como analisar e descobrir relacionamentos entre os mesmos.

Inteligência Artificial

A Inteligência artificial é a disciplina que tenta compreender os mecanismos através dos quais os seres vivos se baseiam e que permitem que estes possam apresentar um comportamento dito inteligente. A Inteligência Artificial pode ser aplicada sobre dispositivos e equipamentos e equipamentos de modo que estes pareçam exibir comportamentos inteligentes. [Ramos,1998]

A inteligência artificial aplicada aos outros campos caracteriza-se por trabalhar sobre o conhecimento.

Aprendizagem automática

Sua função é fazer com que os dados sejam "aprendidos" pelos programas, tornando-os capazes de tomarem decisões através de "pensamentos", que se fazem mais "inteligentes" cada vez que são acrescentados mais heurísticas ou dados ao seu processamento.

3.2.3 Tipos de Relacionamento entre Informações

Quando se implementam as técnicas de *Data Mining*, procura-se extrair informação para atingir determinados objectivos. São vários os tipos de informação que podemos obter com as técnicas de *Data Mining*. Isto não quer

dizer que só podemos obtê-las através destas técnicas de *Data Mining*, mas com elas conseguimos atingir maior agilidade e abrangência na análise destes.

O dois principais objectivos das técnicas de *data mining* são a previsão e a descrição. A previsão envolve a utilização de variáveis ou campos da base de dados para prever valores futuros ou outras variáveis de interesse. A descrição procura encontrar padrões que descrevam os dados. A previsão e a descrição utilizam as seguintes técnicas de *data mining*.

Classificação

Classificação é uma função de aprendizagem que classifica os dados em uma ou mais classes pré- definidas.

Como a figura 4 mostra pode-se partir os dados em duas classes, apesar de esta não ter uma fronteira que os separe perfeitamente. Esta divisão permite ter numa base de decisão que pode ser aplicada a casos futuros que se enquadrem na mesma situação.

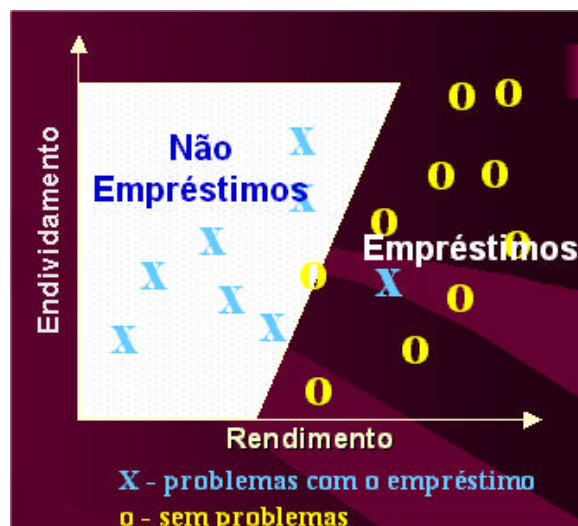


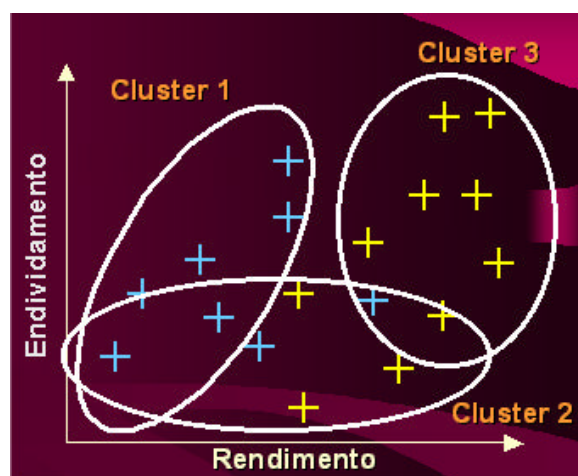
FIGURA 5: CLASSIFICAÇÃO DE UM CONJUNTO DE DADOS REGRESSÃO

A regressão é uma função de aprendizagem que relaciona os dados com uma determinada variável predictiva de valor real. As aplicações da regressão são inúmeras desde de estimação de probabilidades à previsão da procura de um novo produto pelos consumidores após uma campanha de *marketing*.

Clustering

O clustering é uma tarefa descritiva comum que tem por objectivo identificar um conjunto finito de categorias ou agrupamentos para descrever os dados.

As categorias podem ser mutuamente exclusivas e exaustivas, ou consistir numa representação rica tais como, categorias hierárquicas ou sobrepostas. Como é mostrado na Figura 5 existem 3 clusters que se interceptam entre si, havendo dados que pertence a mais do que um cluster.



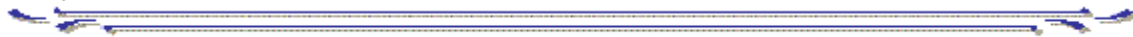


FIGURA 6: CLUSTERS

Sumarização

Envolve métodos de procura descrições compactas para um subconjunto de dados. Os métodos mais sofisticados envolvem a derivação de regras de sumarização, técnicas de visualização de multivariáveis e a descoberta de relações funcionais entre variáveis. As técnicas de sumarização são aplicadas a exploração interactiva dos dados da análise e a geração automática de relatórios.

Modelos de Dependência

Encontrar um modelo que descreva dependências significativas entre as variáveis. Existem dois tipos de modelos de dependência: o nível estrutural e o nível quantitativo. O nível estrutural do modelo especifica quais as variáveis que são localmente dependentes umas das outras (geralmente de forma gráfica). O nível quantitativo do modelo especifica os pesos das dependências usando uma escala numérica.

Alterações e Desvios

Descobrir desvios e anomalias mais significativas nos dados a partir de valores definidos.

3.2.4 Modelação

Surge então a pergunta: Como as técnicas de *Data Mining* é possível achar padrões/relacionamentos entre os dados de uma grande base de dados? Isto é possível através da Modelação.

A modelação é a acção de construir um modelo numa situação onde se sabe a resposta e aplicá-lo a outra situação onde a resposta é desconhecida, analisando as diferenças.

Aplicando-se técnicas para executar a modelação é possível achar padrões relevantes e de interesse do utilizador.

3.2.5 As técnicas e os algoritmos mais utilizados em Data Mining

Árvores de Decisão

Nas árvores de decisão os valores que são encontrados com forte associação são os prognósticos chave ou factores explicativos, normalmente chamados de regras sobre o dados.

As árvores de decisão são meios de representar resultados na forma de árvore, conforme a figura 5, e que lembram um gráfico organizacional horizontal.



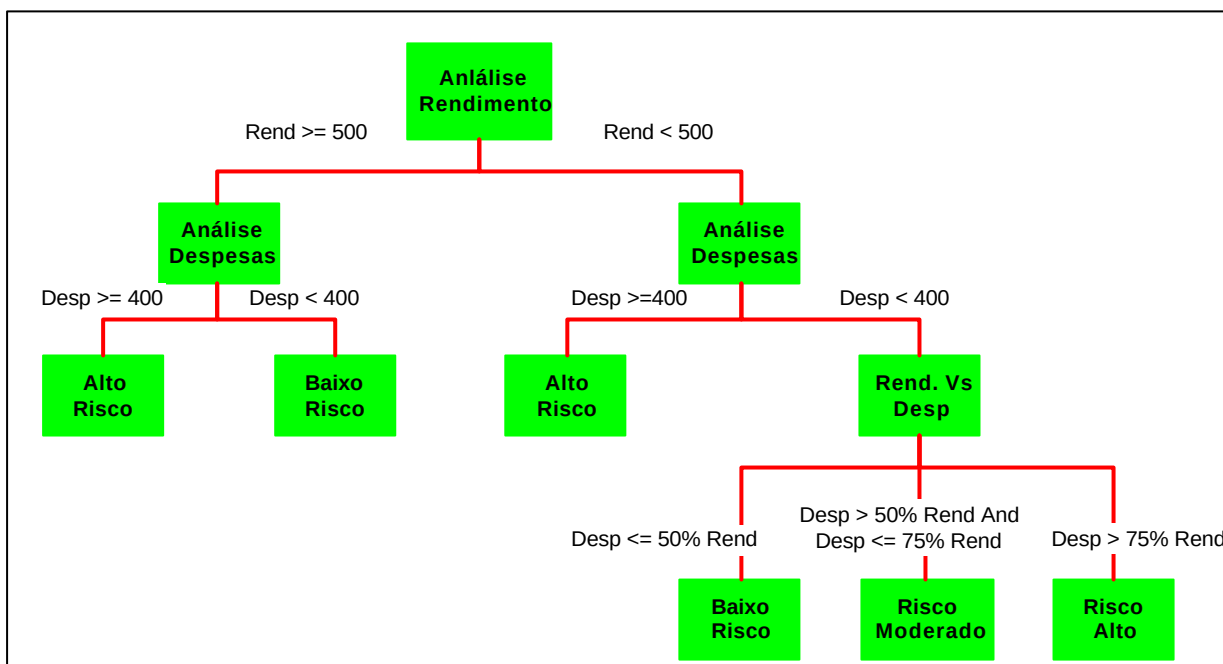


FIGURA 7: ESTRUTURA DA ÁRVORE DE DECISÃO

Possuindo um grupo de dados, com numerosas colunas e linhas, uma ferramenta de árvore de decisão pede ao utilizador para escolher uma das colunas como objecto de saída, e aí mostra o único e mais importante factor correlacionado com aquele objecto de saída como o primeiro ramo nó (raiz) da árvore de decisão. Os outros factores são subsequentemente classificados como subnós dos nós anteriores. Isso significa que o utilizador pode rapidamente ver qual o factor que mais direcciona o seu objecto de saída, e o utilizador pode entender porque o factor foi escolhido. Uma boa ferramenta baseada em Árvore de Decisão permite que o utilizador explore a árvore de acordo com a sua vontade, do mesmo modo que ele poderá encontrar grupos alvo que lhe interessem mais, e aí ampliar o dado exacto associado ao seu grupo alvo. Os utilizadores podem, também, seleccionar os dados fundamentais em qualquer nó da árvore, movendo-o para dentro de uma ferramenta para análise posterior.

As árvores de decisão são, quase sempre, usadas em conjunto com a tecnologia de Indução de Regras, mas são muito importantes, no sentido de apresentar os resultados da Indução de Regras num formato com prioridade. A regra mais importante é apresentada na árvore, como nó raiz, e as regras menos relevantes são mostradas nos nós subsequentes.

Exemplos de ferramentas: Alice d'Isoft, HyperParallel, Discovery, Business Objects BusinessMiner, DataMind, Angoss Knowledge Seeker.

Indução de Regras

A Indução de Regras, rege-se como uma lista, apresentada aos utilizadores através da detecção de tendências dentro de grupos de dados, ou de "regras" sobre o dado, com pouca interacção do utilizador;

O analista de negócio neste caso tem que fazer a ordenação das regras por importância, determinando aquelas que são mais importantes no modelo de *Data Mining*, e para os assuntos de negócios envolvidos.

Exemplos de ferramentas: IDIS, Knowledge Seeker;

Redes Neurais

As redes neurais oferecem o mais profundo poder de exploração dos dados. Porém é a técnica mais difícil de ser entendida pelo utilizador final. As apresentações internas, de como os padrões e os modelos foram achados nos dados, não são mostrados. Estas "caixas pretas" são inexplicáveis, o que faz desta técnica inaceitável para alguns analistas de negócios.

Utilizam um modelo idêntico ao dos neurónios e suas ligações (sinapses) do cérebro humano, como suporte para trabalharem com um conjunto de dados. Aprendem a partir do treino sobre esses dados, generalizando padrões, aplicando-os internamente para classificação e previsão de resultados, conforme a figura 6.

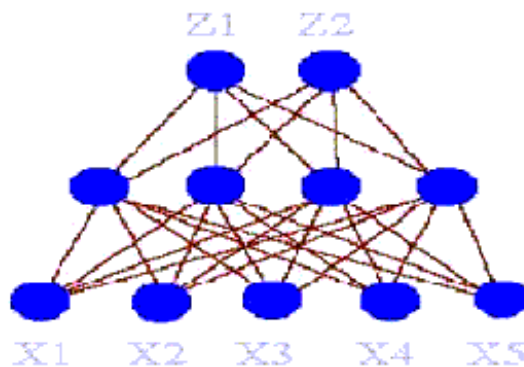


FIGURA 8: ESTRUTURA DE UMA REDE NEURONAL

Cada neurónio tem geralmente um conjunto de pesos que determina como o neurónio avalia a combinação dos sinais de entrada. A entrada para um neurónio pode ser positiva ou negativa. A aprendizagem faz-se pela modificação dos pesos usados pelo neurónio de acordo com a classificação de erros que foi feita pela rede como um todo. As entradas são geralmente pesadas e normalizadas para produzir um procedimento suave.

Durante a fase de treino, a rede estabelece os pesos que determinam o comportamento da camada intermediária. O termo propagação retroactiva (*back-propagation*) é usado quando os pesos são ajustados baseados nas estimativas feitas pela rede - suposições incorrectas reduzem os limites para as conexões apropriadas.

É uma técnica excelente, muito vantajosa, mas que tem dois aspectos negativos. O primeiro é a complexidade dos modelos gerados. O segundo é o problema de exigir que o formato dos dados seja imutável, para diferentes representações dos dados surgem resultados diferentes.

Exemplos de ferramentas: SPSS Neural Connection, IBM Neural Network Utility, NeuralWare NeuralWorks Predict

Regras de associação (*Market Basket Analysis*)



Técnica de localização que permite através de um cesto de compras, inferir regras de associação a partir de grupos de itens ou produtos que surgem nesta transacção.

A análise combinatória dos produtos juntamente com a localização geográfica do cliente, permite promover investimentos em marketing específicos para determinados produtos e regiões.

Para vendas não anónimas é possível ainda promover investimentos em marketing para determinados perfis de clientes, focando os seus traços de compra, garantindo maior retorno no investimento.

Raciocínio Baseado em Memória

Usa relações já conhecidas como *templates*, para inferir e fazer previsões de relações a serem descobertas. A fiabilidade da opção dependerá da distância a que estiverem os alvos vizinhos.

A sua grande vantagem é que pode ser utilizada com inúmeros tipos de dados. Permite ainda aprender novas classificações apenas com a introdução de novas instâncias na base de dados.

On-Line Analytical Processing (OLAP)

Mesmo não sendo uma técnica de *Data Mining* é excelente como ferramenta para a visualização da informação das bases de dados relacionais, de um modo agradável para o utilizador. Serve também para a pesquisa de padrões importantes na informação.

É considerada uma das funções mais sofisticadas em Análise de Dados, pois permite e possibilita aos utilizadores estudarem os dados de maneira multidimensional, podendo perfurar os dados até os seus detalhes (*drill-down*), ou ver porções sumariadas destes dados (*slice-and-dice*).

On-Line Analytical Processing usa um modo especial de modelar os dados do Data Warehouse/Mart, os chamados Cubos-OLAP. Proporciona as condições de análise de dados para responder às possíveis perguntas dos analistas, gerentes e executivos.

A funcionalidade é caracterizada pela análise multidimensional e dinâmica dos dados. O próprio utilizador gera consultas de acordo com suas necessidades de cruzar as informações com métodos que o levem a descoberta daquilo que procura com simples clicks do rato. Mas, relatórios OLAP limita-se a perguntas o que aconteceu?, já o *Data Mining* explica o porquê das coisas, proporcionando uma informação proactiva.

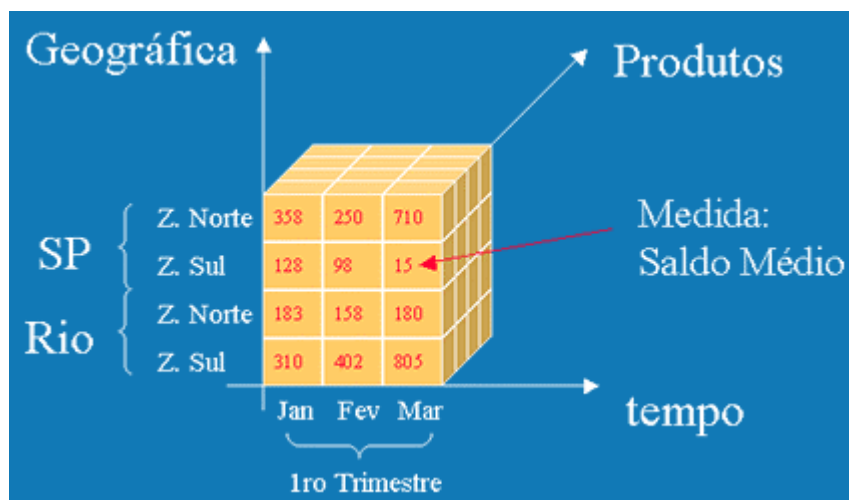


FIGURA 9: OLAP

Drill Down / up - É o processo de analisar a informação em diferentes níveis de detalhes. Por exemplo, uma fotografia do planeta tirada por satélite. Esta fotografia mostra o planeta terra, com seus oceanos e continentes. O utilizador acede à fotografia e pede uma visão mais aproximada do continente sul-americano. A fotografia amplia mostrando esse detalhe e assim por diante. Outro exemplo é a visualização das estatísticas de vendas de um determinado produto em um determinado ano. Esta informação pode ser ampliada para vendas por mês durante esse ano ou por dias. Permite também uma visão mais global como a das vendas na década, produtos mais vendidos em determinada região etc.

Roll-UP - A operação de *roll-up* é o oposto do *Drill Down /up*, já que fornece uma visão dos dados cada vez mais generalizada.

Slice-and-dice - Significa a redução da amplitude dos dados. Por exemplo, podemos com esta característica, analisar as vendas apenas do produto A, apenas na região B.

Exemplos de ferramentas: construtores e analisadores de armazéns de dados

DATA MART

Os Data Marts possuem a mesma estrutura dos Armazéns de dados. O que os difere é a sua amplitude. Enquanto os Armazéns de dados englobam toda a empresa, os Data Marts são projectados para atender as necessidades de um determinado departamento, área da empresa ou grupo de utilizadores.

Os Data Marts podem ser usados independentemente, integrados noutros Data Marts ou ainda interligados para, juntos construírem um Armazém de dados.

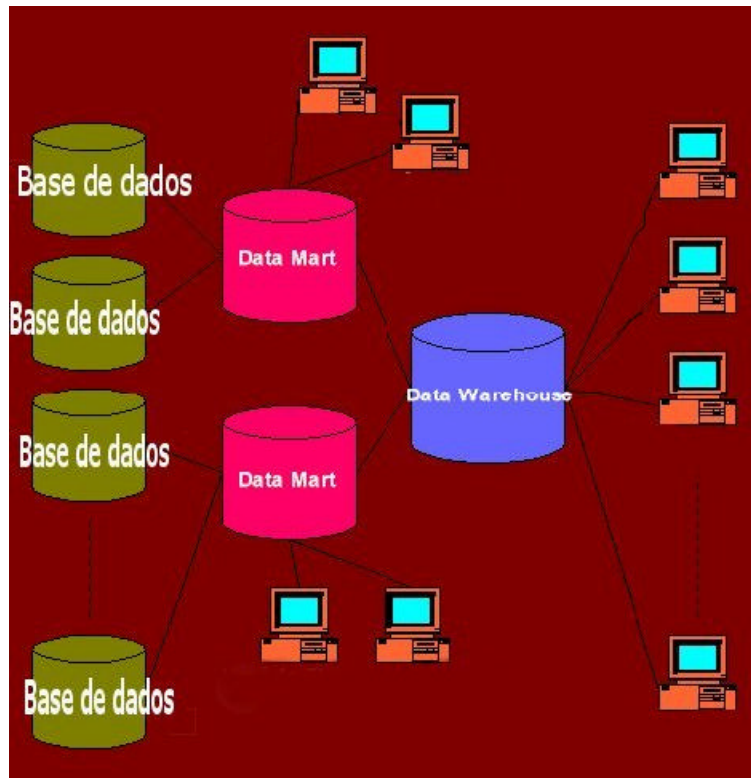


FIGURA 10: DATAMART

Arquitectura

Existem duas formas de se criar um Data Mart. A primeira é denominada *top-down* e a segunda *bottom-up*. No *Top-down*, a empresa cria um Armazém de dados. Depois divide este em áreas menores, gerando assim pequenas bases de dados, orientados por assuntos ou departamentos. No *Bottom-up* a situação é inversa. A empresa opta por criar primeiro um base de dados para um departamento ou grupo de utilizadores. Com isso os custos são bem inferiores de um projecto de Armazenagem de dados completo.

3.2.6 Abordagens de utilização as técnicas de Data Mining

As ferramentas de *Data Mining*, podem trabalhar em cima de qualquer agrupamento estruturado de dados, não precisa necessariamente de ser uma base de dados formal. Isto elimina a obrigatoriedade de aplicar estas ferramentas em Armazéns de dados. Deve-se porém realçar que, para ter maior êxito na descoberta de associações nos dados ou ainda transformá-los como base para tomada de decisão, o mais indicado é ter estas ferramentas "a trabalhar dentro" de dados bem organizados e abrangentes, que é um dos objectivos ao criarmos armazéns de dados ou DataMarts.

Áreas de uso

As três grandes áreas de uso de um *Data Mining*, de modo simplificado, são:

- **Cruzamento de vendas (Cross-selling):** Identificar associação entre produtos, procurando tirar proveito dessas associações. "Descobriu-se numa determinada cadeia de lojas que ao comprar um brinquedo, geralmente eram também compradas barras de chocolate";
- **Aumento das vendas (Up-selling):** Identificar entre os clientes actuais aqueles que tem potencial para adquirir produtos que proporcionem maior lucro para a companhia. "Através da definição de um perfil de utilizador, que compra um produto, podemos determinar que outros utilizadores, do mesmo perfil, e que ainda não tem este produto, venham a comprá-lo";
- **Fidelização:** Combinar e descobrir factores que causem a perda de clientes e como forma de economia, manter estes clientes, o que é mais barato que conquistar novos.

Descrevendo alguns exemplos em determinadas áreas de aplicação, é de notar o poder das técnicas *Data Mining*, e como elas podem determinar a melhor escolha. Nota-se que as técnicas de *Data Mining* apesar de se tratar duma tecnologia recente, está presente no dia-a-dia:

Vendas (Marketing)



✗ Com a análise de compras de seus clientes através dos seus cartões, a *American Express* oferece promoções e propagandas particulares, para os seus clientes.

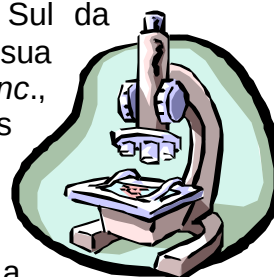
✗ A SONAE através de seu Cartão "Universo", usado nas suas lojas e hipermercados, guarda referências de todas as compras efectuadas pelos respectivos clientes, direccionando o atendimento e promoções.

A utilização das técnicas de *Data Mining*, veio favorecer a área de marketing, pois consegue-se particularizar e criar perfis de consumo, e com isto criam uma ligação mais próxima entre a empresa e o cliente.

Os ganhos desta área tornam-se mais evidentes quando o direccionamento das ferramentas de *Data Mining* é sectorial dentro de uma organização, e quando esta a auxilia em abordagens JAD (*Joint Application Development*), que consiste na evolução a partir de um protótipo, podendo ser alterada rapidamente, descobrindo tendências ou erros futuros.

Saúde

✗ O Hospital de Problemas de Coluna Vertebral do Sul da Califórnia em *Los Angeles*, pesquisa através da sua ferramenta *Data Mining*, o IDIS da *IntelligenceWare, Inc.*, factores subtis que afectem o sucesso ou o insucesso das intervenções cirúrgicas à coluna vertebral.



Esta ferramenta está a ser utilizada também pela *U.S. Gymnastics Federation*, em pesquisa de factores que, a longo prazo, contribuam para o rendimento de um atleta.

Ainda nos EUA, o IDIS é utilizado para pesquisas e abordagens nos registos médicos em factores que contribuam ou que tenham relação no tratamento do cancro.

✗ A redução de custos obtidas pelo Health-KEFIR nos cuidados com a saúde e o simultâneo melhoramento desses mesmos cuidados, parece contraditório, porém segundo *Gregory Piatetsky-Shapiro*, investigador principal no *Knowledge Discovery in Databases Project* nos *GTE Laboratories* - "Os relatórios gerados pelo KEFIR são mais completos, gerados em menos tempo, e significativamente mais baratos do que relatórios idênticos de uma consulta médica tradicional", isto após a implantação das ferramentas de armazém de dados e *Data Mining*.

Com este sistema é possível seleccionar apenas condições interessantes de saúde, como por exemplo: mulheres com gravidez normal não são seleccionadas, enquanto que as que tiverem nascimentos prematuros são, canalizando os esforços para intervenções médicas (de rotina) que promovam a redução da taxa de nascimentos prematuros, evitando as despesas maiores dos cuidados excepcionais necessários para este tipo de ocorrência.

Ciência e Tecnologia

✗ Foram descobertos recentemente nove quasares, através da ferramenta de *Data Mining* SKICAT. Seriam necessários três anos para descobrir um número idêntico de quasares de mesmo tipo, utilizando as técnicas tradicionais de pesquisa.



Com o SKICAT, os astrónomos de Caltech conseguiram o mesmo feito em menos de seis meses.

O uso de *Data Mining* juntamente com os armazéns de dados na Ciência, acelera a descoberta de novos paradigmas. O tempo para descobrir relações detectáveis em ambientes vastos estão a ser drasticamente reduzidos, ocasionando evoluções mais rápidas.

Finanças

O foco principal de utilização das ferramentas de *Data Mining* na área financeira, é a prevenção e detecção de fraudes.

✗ O *Security Pacific/Bank of America* utiliza ferramentas de *Data Mining* no suporte à decisão na área dos empréstimos bancários, com intuito de prevenir fraudes.

✗ O BPP (Banco Privado Português) sentiu a necessidade de um suporte à decisão de avaliação de perfis de riscos para os investimentos financeiros dos seus clientes. Com a implantação de ferramentas de *Data Mining*, foi possível criar modelos devidamente enquadrados nos segmentos alvo de cada cliente.



Através de perguntas envolvendo característica sócio-económicas, risco, objectivos, etc, foi aplicado um conjunto de técnicas de aprendizagem a um universo de clientes definidos pelo Banco. Deste processo resultou a definição de um modelo capaz de identificar as perguntas mais relevantes a cada instante, no sentido de conseguir uma caracterização adequada ao perfil de risco do cliente. Além da estratégia de investimento, o modelo devolve também uma primeira análise da distribuição das respostas dadas, permitindo a identificação de incoerências nas respostas e de eventuais pontos que devem ser esclarecidos para assegurar que o cliente está perfeitamente identificado com a estratégia proposta.

Propostas de Uso

✖ Com a devida organização dos dados, as ferramentas de *Data Mining* podem ser capazes de analisar informações de regiões/estados, nas mais diferentes instâncias: cultura, distribuição de renda, senso demográfico, etc.

Tomando-se como base que o país seria como uma grande empresa, poderíamos ter um Data Mart estatal, e através da identificação de relações entre estes e as ferramentas de *Data Mining*, seria possível agregar e planejar estratégias de política comuns, aprender com erros ocorridos numa área e aplicar soluções correctas de um estado para outro, além de poder auxiliar o controlo e libertação de gastos públicos.

✖ No meio publicitário seria possível descobrir relações entre notícias oficiais, entre as mais diversas esferas, tendo catalogados todos os ofícios, memorandos e decisões políticas. E com as ferramentas de *Data Mining*, cruzar informações e prever situações que ocorrem devido/após o lançamento de um anúncio oficial.

Há registo do uso efectivo de *Data Mining* em países como os EUA, a França ou a Alemanha, onde estas ferramentas já estão presentes em algumas decisões, tomadas pelos seus governantes.

3.2.7 Dificuldades para o uso de Data Mining

Existem alguns impedimentos na utilização das técnicas de *Data Mining*. O mais significativo hoje ainda é o alto custo, porém vários factores podem impedir sua implantação:

Organização

De nada adianta investir nas técnicas de *Data Mining* se os dados da empresa estiverem espalhados, redundantes ou se não forem fiáveis. Quando analisamos dados referentes a uma transacção estes não podem ser contraditórios.

Geralmente a implantação de armazéns de dados precede a implantação de ferramentas de *Data Mining*, pois possibilitam agregar informações espalhadas de diferentes estruturas, formas e plataformas. Para tal, a definição e organização destes dados devem ser minuciosas, uma vez que os dados mal organizados afectam negativamente os resultados da análise.

Alto Custo

Este é sem dúvida um dos principais motivos de desistência na implantação de soluções de *Data Mining*. O alto custo da maioria das ferramentas dificulta a disseminação desta ferramenta entre as empresas. Alguns fornecedores têm introduzido produtos com custo mais baixo mas, mesmo assim, o preço continua a ser a principal limitação da sua utilização em larga escala.

A tendência é que estes preços caiam, devido ao amadurecimento e divulgação da tecnologia.

Grandes Bases de Dados

O maior obstáculo às técnicas de *Data Mining* no passado foi a necessidade de armazenar e administrar grandes quantidades de dados, e de servidores. Isso por si só já dificultava bastante o crescimento no mercado das técnicas de *Data Mining*. No entanto, a maioria dos fornecedores dessa tecnologia continua a insistir no discurso de que as ferramentas de *Data Mining* requerem terabytes de dados e poderosos servidores, mas existem soluções mais acessíveis no mercado.

Novidade e Complexidade

Devido a ser uma nova geração de ferramentas, para a extracção de conhecimento em grandes conjuntos de dados uma terceira barreira ainda permanece: a grande maioria das ferramentas ainda continua incompreensível para os utilizadores comuns. De facto, muitas ferramentas ainda fazem o seu trabalho numa "caixa-preta", não permitindo que se saiba como alcançaram os seus resultados.

Isso significa que a utilização da técnicas de *Data Mining* ainda tem que ser feito, no contexto da área de sistemas a quem os utilizadores têm que submeter as suas solicitações, esperar por dias ou semanas enquanto um perito processa os dados, para então receberem e examinarem a saída consolidada. Se os resultados não forem satisfatórios, todo o processo tem que ser recommçado. Já existem porém soluções mais compreensíveis e a tendência é que se aproximem ainda mais dos utilizadores.

Montar um Ambiente Ideal

Estruturar um ambiente ideal, conforme a figura 9, é uma tarefa árdua, devido aos anos e anos de desorganização na forma de guardar as informações nas instituições.

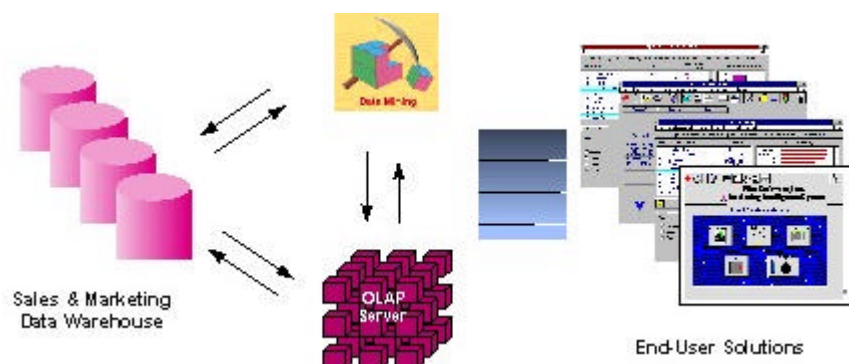


FIGURA 11: AMBIENTE IDEAL PARA UMA FERRAMENTA DE *DATA MINING*

Preparar os dados para utilizar ferramentas de *Data Mining*, é considerado como sendo 80% do trabalho geral na utilização de uma ferramenta de *Data Mining*, por englobar inúmeras e trabalhosas tarefas.

O ambiente ideal deve conter dados relevantes para as necessidades dos utilizadores, limpos (livres de erros lógicos ou de entrada de dados), consistentes, e sem falta de dados.

3.3 Text Mining



Para tomar as decisões correctas é preciso analisar muitas vezes um grande volume de informação textual. O *text mining* permite a extracção do conhecimento em dados não estruturados: texto. Sendo assim o *text mining* faz uma análise quantitativa de grandes volumes de textos ajudando a compreender todo o conhecimento disponível, que seria impossível obter de forma manual.

Pode ser aplicado a uma grande variedade de documentos, memorandos, cartas, contratos, patentes, transcrição de discursos, mensagens de correio electrónico, e outras fontes similares.

Serve para gerir conhecimento, correspondendo à experiência colectiva, conhecimento e sabedoria da organização.

O conhecimento, para além da informação propriamente dita, engloba o contexto, alternativas e factos dos processos de decisão, que se podem encontrar nos documentos da organização

Apesar de o *text mining* ter várias analogias com o *data mining* tradicional, diferem na medida em que o último funciona essencialmente através da aplicação de algoritmos de aprendizagem e estatísticas a conjuntos de dados, identificando relações previamente desconhecidas, ao passo que o *text mining* refere-se ao processo de procura através de documentos não estruturados – apesar de muitos documentos possuírem campos bem definidos como título, autores, categoria, entre outros, o corpo do documento costuma não ter uma estrutura rígida – inferindo algum significado deles.

Benefícios do *text mining*

- aumentar o valor da *informação* da empresa;
- baixos custos de integração *versus* outras técnicas de processamento texto;
- produtividade aumentada dos trabalhadores que lidam com o conhecimento;
- competitividade melhorada.

Text Mining - associação por palavras-chave

Pegar num documento e construir uma tabela com as palavras mais relevantes desse mesmo documento, para depois tentar encontrar associações entre essas palavras. Ex. [Universidade, Coimbra], [Portugal, Fado, Amália].

Text Mining - classificação de documentos

O *test mining* permite organizar (automaticamente) grandes quantidades de documentos por classes para posteriormente serem utilizados em pesquisa documental.

É construído um conjunto de documentos pré-classificados para conjunto de treino, sendo este conjunto analisado de modo a gerar modelos de classificação. Estes modelos são geralmente refinados através de processos de testes. O modelo assim gerado é utilizado para a classificação de outros documentos.

Um modo eficaz de classificação de documentos é explorar o método anteriormente referido (classificação por associação), classificando os documentos da seguinte forma:

- extracção de palavras-chave e termos por pesquisa de informação e outras técnicas de associação;
- criação de hierarquias de palavras-chave e termos, usando classes de termos, ou outra base de conhecimento.
- depois de definido este modelo, trata-se de aplica-lo para a classificação de outros conjuntos de documentos.

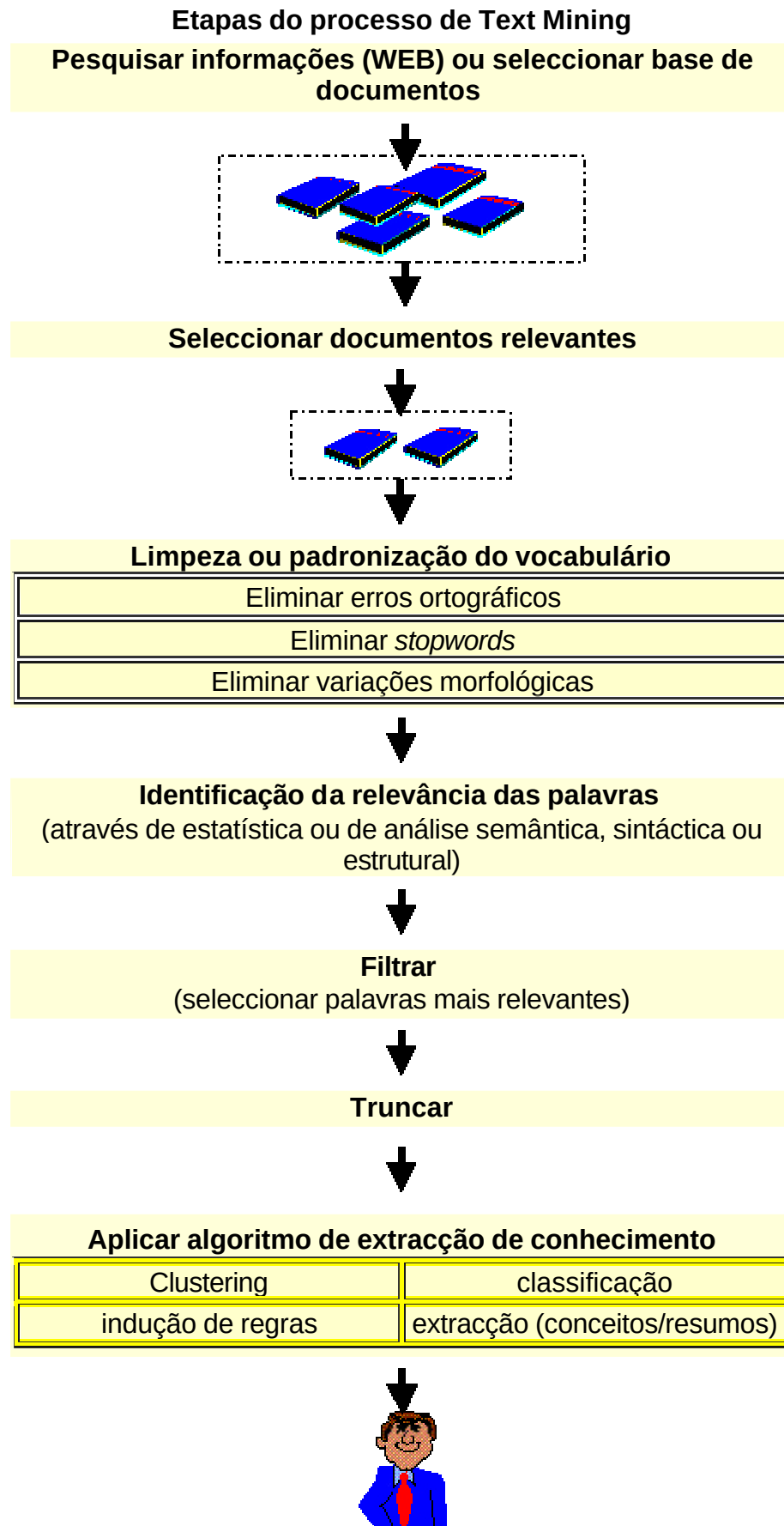
O processo de classificação de documentos é também bastante eficaz para a classificação de páginas *Web*.

Outras aplicações de text mining

- Execução de sumários - consiste em sumariar textos, tentando apanhar as ideias mais importantes;
- Clustering - análise de diversos documentos, e agrupamento de pedaços (*clusters*) de texto de acordo com semelhanças no conteúdo;
- Identificação de línguas - consiste em descobrir em que linguagem determinado documento está escrito
- Detecção de duplicados em arquivos

Exemplos

- gestão de correio electrónico;
- gestão de documentos ;
- *help desk* automatizado;
- pesquisa/investigação de mercado;
- recolha de “inteligência” de negócio [GIL&RODIGUES,2001].



3.4 Web Mining

O *Web Mining* tem sido definido como uma aplicação de técnicas de *Data Mining* aos dados da Web. O objectivo do *Web mining* é encontrar algo de interessante no amontoado de informação tanto para o utilizador como para o dono do site. Sendo assim pode-se dividir em dois objectivos. O desenvolvimento de técnicas que ajudem o utilizador a navegar na Web:



Objectivo para o utilizador

- Encontrar documentos de um tema;
- descobrir documentos semanticamente relacionados;
- prevenir a desorientação do utilizador.

E o desenvolvimento de técnicas que ajudam no desenvolvimento de produtos e serviços e no design de sites da Web:

Objectivo do dono do site

- aumentar o contacto com o cliente (*Web Marketing*);
- direccionar promoções de produtos;
- desenvolvimento de serviços personalizados.

Sendo assim, existe a necessidade de técnicas que ajudem as decisões de suporte a negócios e no comércio electrónico. Um exemplo é a necessidade de compreender o comportamento do utilizador para que se possa focar a publicidade na Web. Outro exemplo é o design dos sites que consegue atingir o objectivo de negócio (o site que leva o utilizador às páginas com os produtos que dão mais lucro).

Por outro lado existe uma comunidade inteira que visa dar ao cliente ferramentas e serviços que têm como objectivo melhorar a experiência de navegação do utilizador. Sendo assim o utilizador beneficiará do acesso a motores de pesquisa com técnicas melhoradas, e *Web sites* com interfaces personalizados e com ferramentas pessoais para ajudar a lidar com a grande quantidade de informação e opções de navegação da Web.

A figura seguinte mostra um esquema de acesso à Web às possíveis fontes de informação que se podem tirar deste acesso, fontes estas que vão dar origem aos três campos de estudo do *web mining*.

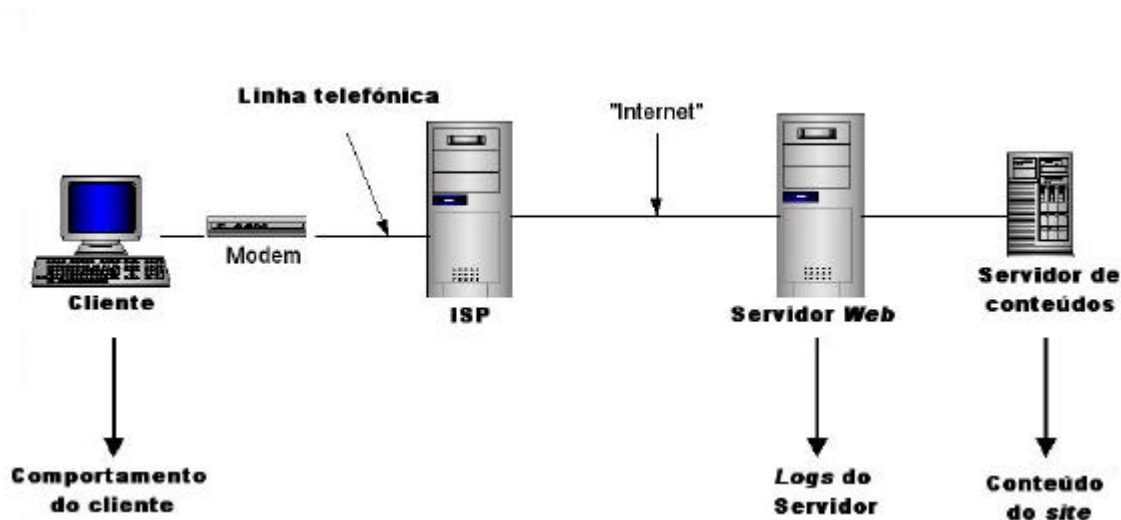


FIGURA 12: ACESSO À WEB

Sendo assim o web mining divide-se em: *Web Usage Mining*, *Web Structure Mining* e *Web Content Mining*. Como é exemplificado pela figura seguinte.

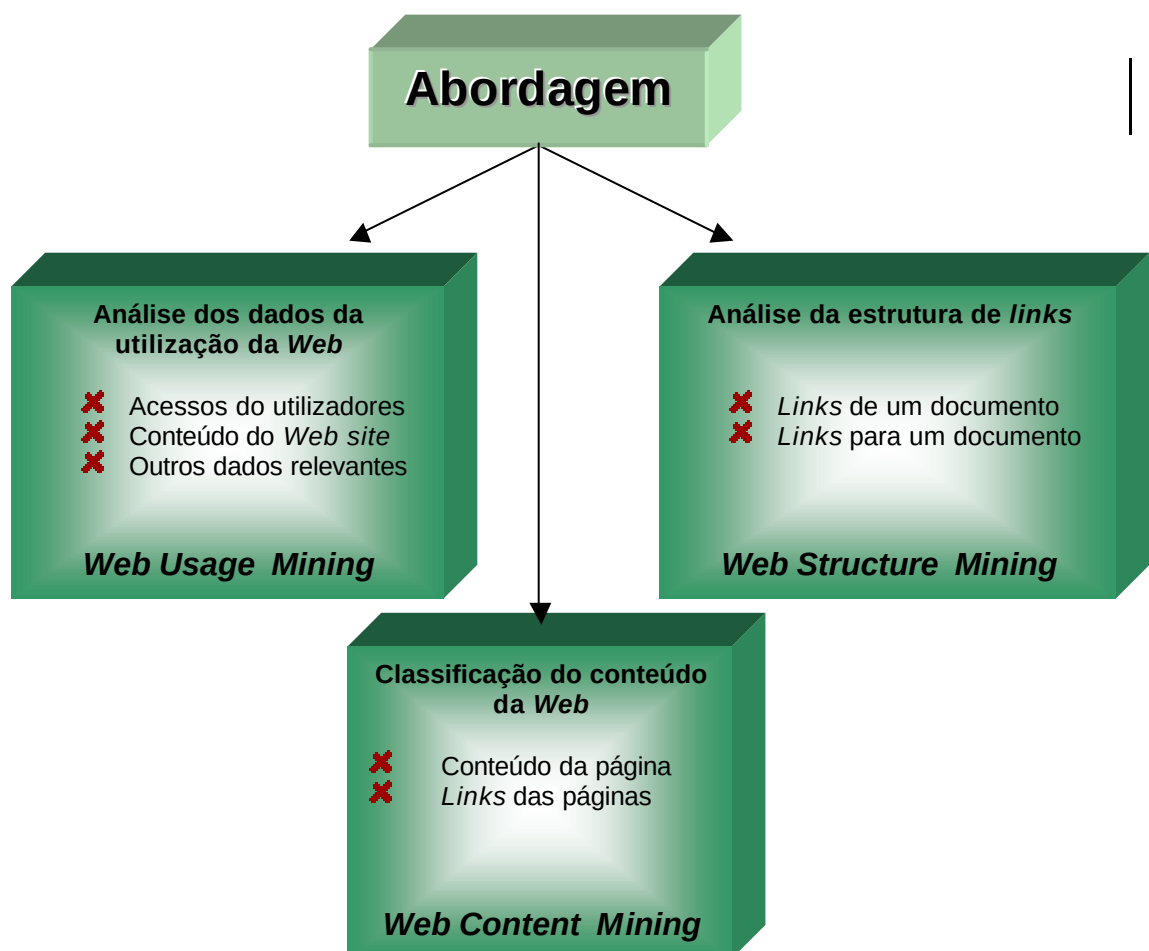


FIGURA 13: ABORDAGENS DO WEB MINING

3.4.1 Web Content Mining

Web content mining refere-se as técnicas que assistem o utilizador a encontrar documentação dentro de um determinado critério. A descoberta da semântica dos documentos pela exploração:

- do conteúdo;
- dos *links* de e para o documento;
- da sabedoria dos peritos;
- das selecções dos utilizadores.

Os documentos encontrados são utilizados para formar uma base local de conhecimento. Uma das abordagens para a solução deste problema é a definição de uma linguagem de interrogação que é baseada nos recursos da *Web*. Uma das técnicas é a utilização de um modelo de base de dados multi-camada que transforma os dados da *Web* não estruturados num formulário de acordo com a tecnologia de base de dados. As ferramentas especializadas são utilizadas para extracção de informação da página *Web* com o objectivo de identificar os documentos relevantes. As características dos documentos são guardados numa base de dados local. A linguagem de interrogação permite a interrogação de um grande número de características na base de dados local e, se mais detalhe for necessário, os recursos actuais da *Web* serão interrogados pelos motores de pesquisa.

3.4.2 Web Structure Mining

World Wide Web pode revelar muito mais informação do que a informação contida nos documentos. Por exemplo, os *links* que apontam para determinado documento indicam a sua preferência enquanto os *links* que saiem de um documento podem indicar a variedade de tópicos abrangida por um documento. Isto pode ser comparado a citações bibliográficas, quando um documento é varias vezes citado pode significar o tamanho da sua importância. O método de *PageRank* tira partido desta informação contida nos *links* para encontrar as páginas mais citadas. Os contadores de *links* num documento reconstroem a estrutura das páginas da *Web*.

O *Web Structure Mining* incide o seu campo de investigação na análise da estrutura de *links* da *Web* para poder identificar documentos relevantes.

O conteúdo dos documentos reflecte-se:

- nos documentos que cita;
- nos documentos em que é citado;
- artigos e livros similares.

PageRank

PageRank é o método pelo qual é medida a importância de uma página. Quando factores, como o título, palavras-chaves, são levados em consideração, então é utilizado o



PageRank para ajustar o resultados dos motores de pesquisa (Ex. *Google*), os *sites* mais importantes são mostrados em primeiro lugar.

O método *PageRank* aproveita as características da *Web*, usando sua vasta estrutura de *links* como uma ferramenta organizacional. Em resumo, interpreta um *link* da página A para a B como um voto, da página A, para a página B. O motor de pesquisa *Google* determina a importância de uma página pelos votos recebidos. No entanto, é preciso algo mais que um volume absoluto de votos, ou *links*; é analisada também a página que lançou o voto. Páginas "importantes" que lançam um voto são de grande peso e ajudam a tornar outras páginas "importantes".

Esses resultados importantes e de alta qualidade recebem uma *PageRank* mais elevada e serão colocados em posições mais altas nos resultados. Sendo assim, a *PageRank* é um indicador geral de importância e não depende de uma consulta específica. Ao contrário, é uma característica da página baseada nos dados da *Web* que o analisa por intermédio de algoritmos complexos que determinam a estrutura do *link*.

Sem dúvida, páginas importantes não significam nada se não corresponderem à consulta efectuada. Por isso, são usadas sofisticadas técnicas de correspondência de texto para localizar páginas que são importantes e relevantes para a pesquisa. Por exemplo, quando se analisa uma página, procura que as páginas que a ela estão vinculadas, dizem a respeito dela.

Qual a ordem de classificação do motor de pesquisa *Google*:

1. Encontra todas as páginas que menciona as palavras chave da pesquisa.
2. São ordenadas de acordo com as palavras chave encontradas.
3. Calcula no interior do texto associado às hiperligação
4. O resultado é ajustado pelo *PageRank*.

A informação de estrutura da *Web* divide-se em:

- ▮ Intra-página
 - ↪ refere-se à estrutura interna do documento *Web* em HTML ou XML, que é normalmente representado através de uma árvore.
- ▮ Inter-páginas
 - ↪ pode ser analisada por travessia dos *links* (estruturas dos *links Web*)
 - ↪ normalmente são usados grafos em que os nós são páginas e os *links* os ramos dos grafo.

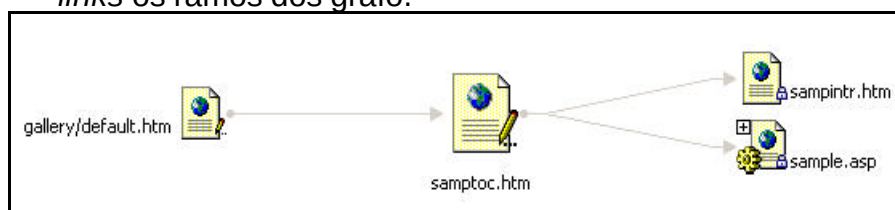


FIGURA 14: ESTRUTURA DE LINKS

3.4.3 Web Usage Mining

Web Usage Mining é um campo de investigação que se baseia no desenvolvimento de técnicas e ferramentas que estudam o comportamento da navegação dos utilizadores. Compreender as preferências de navegação do utilizador é um passo essencial para o estudo da qualidade de um *site* de comércio electrónico. Compreender os padrões de acesso dos utilizadores permite aos fornecedores personalizar e adaptar o interface do *site* para um utilizador.



Quando os utilizadores interagem com o *site*, os dados do seu comportamento são guardados em ficheiros de registo de acessos. Estes ficheiros de registo de acessos podem conter informação que caracteriza a experiência do utilizador no *site*. Como os ficheiros correspondem a muitos megabytes por dia, há uma necessidade de técnicas e ferramentas que permitem tirar partido desse conteúdo.

- o comportamento do utilizador é registado;
- a prospecção dos registos diz como o *site* é compreendido;
- adaptação do *site* aos utilizadores;

Web Usage Mining permite ajustar o *site* à percepção do utilizador levando assim à satisfação do cliente.

O *Web Usage Mining* pode ser de dois tipos:

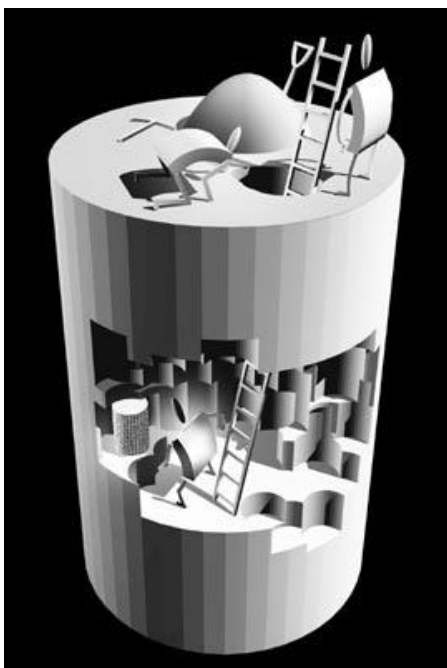
- não Personalizada – o utilizador é observado como uma unidade sem identidade (ex. Análise de cesto de compras);
- personalizada – o utilizador é observado como uma pessoa individual, da qual são sabidas a identificação e dados pessoais (ex. Análise dos dados dos clientes).

Ficheiros de registos de acessos e preparação de dados

O ficheiro de registo de acessos é um ficheiro de texto onde são registados, todos os acessos a uma página do servidor *Web*. Cada pedido corresponde a uma entrada no ficheiro que contém normalmente a seguinte informação:

- ⚡ o endereço IP do computador que fez o pedido;
- ⚡ a identificação do utilizador (este campo não é utilizado na maioria dos casos);
- ⚡ data e hora do pedido;
- ⚡ campo de estado que indica se o utilizador teve sucesso;
- ⚡ o tamanho do ficheiro que foi transferido;
- ⚡ o URL, da página que gerou o pedido;
- ⚡ o nome e versão do *browser* utilizado.

Esta informação pode ser utilizada para reconstituir as sessões de navegação do utilizador, dentro do *site*. Num cenário ideal para cada utilizador seria alocado um único endereço IP, cada vez que o utilizador visita o *site*. É esperado que o utilizador visite o *site* mais do que uma vez e sempre com um objectivo diferente. Sendo assim uma sessão de utilizador é normalmente definida como uma sequência de pedidos vindos do mesmo endereço IP, sendo assim dois pedidos serão separados por X minutos.



No entanto a identificação de uma sequência de pedidos feitos por um dado utilizador, quando este visita um *site*, é uma tarefa mais complicada. Nem todos os pedidos são registados num ficheiro de registo de acessos e um utilizador pode alocar mais de um endereço de IP durante uma única sessão. Sendo assim a etapa de preparação de dados do *Web usage mining* exige um esforço considerável.

A primeira tarefa quando utilizamos um ficheiro de registo de acessos, consiste em identificar todas as entradas válidas. O protocolo de transferência de ficheiros utilizados na *Web* requer o estabelecimento de uma conexão separada para cada ficheiro pedido. Sendo assim, uma página HTML contendo imagens, som ou vídeo, vai originar

um pedido para cada ficheiro que contem. Na maior parte dos casos a entrada do ficheiro de registo de acessos corresponde ao ficheiro pedido pelo utilizador, e todas as outras entradas serão ignoradas.

Outro aspecto relevante a ser tomado em conta quando um ficheiro de registo de acessos é utilizado, é o uso de *cache* e de servidores *proxy* utilizados na *Web*. Como resultado, nem todas as páginas pedidas ao servidor são guardadas no ficheiro de *log*. Se o *browser* encontrar em *cache* uma cópia do

documento que está a ser pedido pelo utilizador, o pedido não será feito ao servidor e a cópia guardada do documento é mostrada. Sendo assim, embora a página seja vista pelo utilizador o pedido não é registado no ficheiro de registo de acessos do servidor.

Uma situação parecida pode ocorrer a nível da *proxy*. O utilizador de *proxies* pode causar dificuldades na identificação dos pedidos feitos por um computador. Um servidor de *proxy* pode ser configurado de tal maneira, que quando é pedida uma cópia de uma página que não está disponível na memória local, esta é pedida pela *proxy*, em nome do utilizador. Nestes casos, o endereço IP guardado no ficheiro de registo de acessos corresponde ao da *proxy* e não o do utilizador.



Note-se que mais do que um utilizador pode utilizar a mesma *proxy* para ver o mesmo site ao mesmo tempo. E é possível ter dois utilizadores a partilhar o mesmo endereço IP, e o mesmo fornecedor de *Internet* pode alocar dinamicamente endereços IP diferentes para o mesmo utilizador durante uma conexão.

Existem técnicas que ajudam a superar estes problemas. Por exemplo, o uso de um campo em conjunção com a topologia do *website* que permite identificar os pedidos que faltam na sessão (por causa do uso da *cache*). Duas sessões separadas, podem permitir identificar as duas sessões simultâneas, mas com um endereço IP comum. O uso da entrada no ficheiro de registos de acessos que identifica o *browser* é útil para distinguir utilizadores com *browsers* diferentes.

O uso de *cookies* permite identificar um utilizador individual dentro do *site*. Se as *cookies* estão activas, quando o documento é pedido por um novo utilizador a resposta inclui um identificador único, o qual é guardado no disco do utilizador pelo *browser*. Todos os subsequentes pedidos feitos pelo *browser* para o mesmo *site* vão incluir a informação da *cookie* e, por isso, permitem que o fornecedor de serviços reconheça o utilizador. No entanto o uso de *cookies* só é possível quando o utilizador consente o seu uso, o que levanta problemas de privacidade. É de notar que a *cookie* apenas identifica *browsers* e não utilizadores individuais.

Outra técnica utilizada por lojas *online*, consiste em adicionar um identificador único à URL da primeira página acedida dentro do *site*. Todas as URLs na página acedida são modificados de forma a incluir o identificador único. Assim, cada URL pedida vai identificar a sua origem, permitindo a um fornecedor de serviços seguir os passos do utilizador dentro do *site*. Se o utilizador se identifica durante o processo de pagamento ele permite ao fornecedor de serviços relacionar a sessão corrente com as sessões anteriores do mesmo utilizador. As sessões de navegação do utilizador podem ser reconstituídas com esta técnica e os problemas de *proxy* e de *cache* ficam diminuídos.

Num cenário em que *browser*, por meio da *proxy* local, guarda num ficheiro de registo de acessos pessoal a caracterização dos dados da navegação da *Web* de um dado utilizador, estes dados podem ser guardados correctamente se a *cache* do *browser* estiver desactivada.

Em conclusão, o relato exacto das sessões de navegação dos utilizadores podem ser afectadas pelo uso de *cache*, servidores *proxy* e partilha de endereços IP. Sendo assim, técnicas como as *cookies* e a modificação de URLs, são essenciais na identificação de pedidos do mesmo utilizador, de forma que se possa fazer correctamente a reconstrução das sessões de navegação de um utilizador a partir do ficheiro de registo de acessos.

Técnicas de Web Usage Mining

Hoje em dia existem diversas ferramentas de análise comercial de ficheiros de registo de acessos. No entanto estas ferramentas têm capacidade limitada de análise, produzindo apenas resultados como estatísticas sumárias e o número de visitas frequentes às páginas. Existem diversas técnicas que permitem tirar partido da informação disponível nos ficheiros de registo de acessos.

Previsão, Personalização e Visualização

É um modelo construído com base na informação passada de utilização e na probabilidade de transições entre as páginas que são proporcionais ao número de vezes que ambas as páginas são acedidas num intervalo de tempo predefinido. Note-se que o uso do intervalo de tempo resulta em transições de probabilidade maior que zero entre as páginas que nunca foram acedidas construtivamente. Os resultados que saem das experiências mostram que o método é eficaz na redução da carga do servidor e do tempo de serviço.

Outro método similar é a dependência gráfica que é deduzida e dinamicamente actualizada enquanto o servidor recebe os pedidos. Existe um nó por cada pedido à página, e um arco entre dois nó, e se o nó alvo é pedido X acessos depois do nó raiz, o peso do arco é proporcional ao número de pedidos. As simulações feitas com os dados do ficheiro de registo de acessos mostram que a redução na recuperação da latência pode ser conseguida. Este método não garante transições, só apenas dentro de pedidos consecutivos.

3.4.4 Linguagens e tecnologias

HTML e XML

Presentemente, a maior parte das páginas disponíveis na *Web* estão escritas em HTML (*HyperText Mark-up Language*). O HTML é uma linguagem que consiste, principalmente, de um conjunto de *tags* que identificam diferentes tipos de componentes numa página (ex. título, tabelas e figuras). Cada componente é posicionado na página de acordo com o seu tipo. O HTML tem sintaxe fixa e limitada em complexidade, é fácil de aprender e relativamente simples de implementar. Estas características permitiram o desenvolvimento da *Internet*, mas quanto mais a *Web* crescia também crescia a ineficácia do HTML.

Com o aparecimento do XML (*eXtensible Mark-up Language*), os pontos fracos do HTML têm sido ultrapassados.

Os três aspectos em que o XML difere do HTML são:

- É uma linguagem extensível, permitindo a definição da estrutura lógica de um documento e impor uma estrutura de validação.
- A adopção do XML permite definir e validar novos tipo de documentos. As definições tipo dos documentos podem ajudar as ferramentas automatizadas a compreender a estrutura lógica dos documentos e permitem um enorme potencial para melhorar a exactidão da tecnologia dos motores de pesquisa.
- XML também melhora a forma como a informação é posicionada no *browser* do utilizador. Por exemplo, permite diferentes formas de ver os componentes de um documento sem ter que fazer *download* de diferentes versões do documento.

XML é uma linguagem de descrição de conteúdos, que permite aos programadores definir as suas próprias marcas e usá-las para anotar os conteúdos disponibilizados nas suas páginas. Utilizando esta linguagem é possível definir, arbitrariamente, uma estrutura para um documento ainda que nenhuma informação exista sobre o significado dessa estrutura.

ASP - Active Server Pages

ASP (*Active Server Pages*) é um ambiente para o desenvolvimento de páginas dinâmicas para a *Web*, que mistura HTML e programação. A extensão destes arquivos é *.asp* e contém combinações de *Server-Side* (códigos de programação que executam no servidor, portanto podem ser acedidos em qualquer *browser*), *scripts* e *tags* HTML.

Alguns recursos que podem ser implementados via ASP:

- Programação em *VBScript* ou *Jscript*;
- Acesso a bases de dados;
- Sessões (persistência de informações no servidor).

As vantagens de se usar ASP

Independência do *browser*

ASP poder executar as páginas complexas no servidor e enviar somente os resultados para o cliente. Todos os *browsers* suportam ASP. Isto acontece pelo facto das páginas ASP serem processadas pelo servidor. O que o cliente recebe é somente código HTML.

Bases de Dados

Permite visualizar, actualizar e adicionar informações nos servidores SQL. Uma aplicação ASP pode ser usada com qualquer base de dados compatível com ODBC. Isto inclui dados do *Access*, *Microsoft SQL Server*, *Oracle*, *Sybase*, *Informix*, *DB2*, entre outros.

Segurança do código fonte

Como o Servidor retorna somente o resultado HTML, o código fonte fica preservado. O recurso ASP é parte integrante do IIS (*Internet Information Server*), que tem sua segurança integrada. É fácil restringir o acesso a páginas ASP usando os processos de autenticação do IIS. É ainda possível dar segurança aos dados transmitidos usando SSL.

Linguagens

O ASP pode utilizar de comandos em *VBScript*, *JavaScript* e HTML. Existem *plug-ins ActiveX* para dar suporte para outras linguagens como *PERL* e *Python*.

Armazenamento de informações

Aplicações ASP podem armazenar dados que são mantidos durante toda uma sessão. Desta forma, um utilizador pode fornecer seu nome somente uma vez numa página e as demais páginas podem obter-se este dado automaticamente. Este recurso é ideal para aplicações de venda pela *Internet*.

ASP EM ACÇÃO

No caso das páginas ASP, o funcionamento é o seguinte: o *browser* do cliente envia a requisição de uma página. O servidor processa o código desta página e retorna ao utilizador uma resposta em HTML.

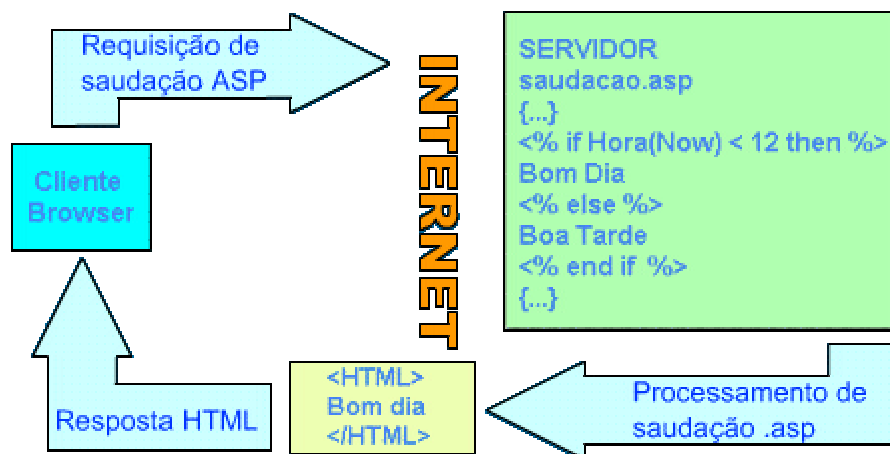
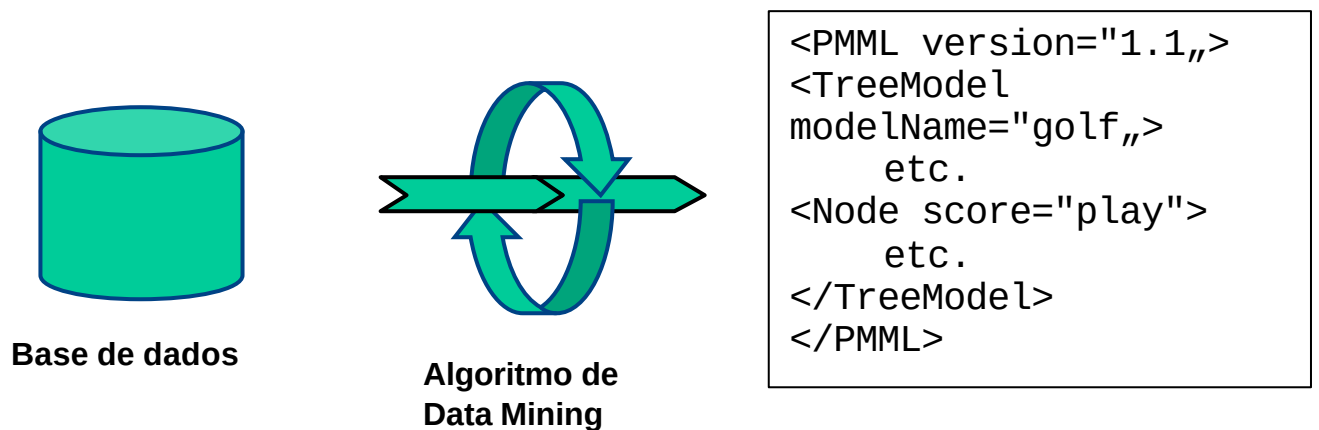


FIGURA 15: FUNCIONAMENTO DAS ASP

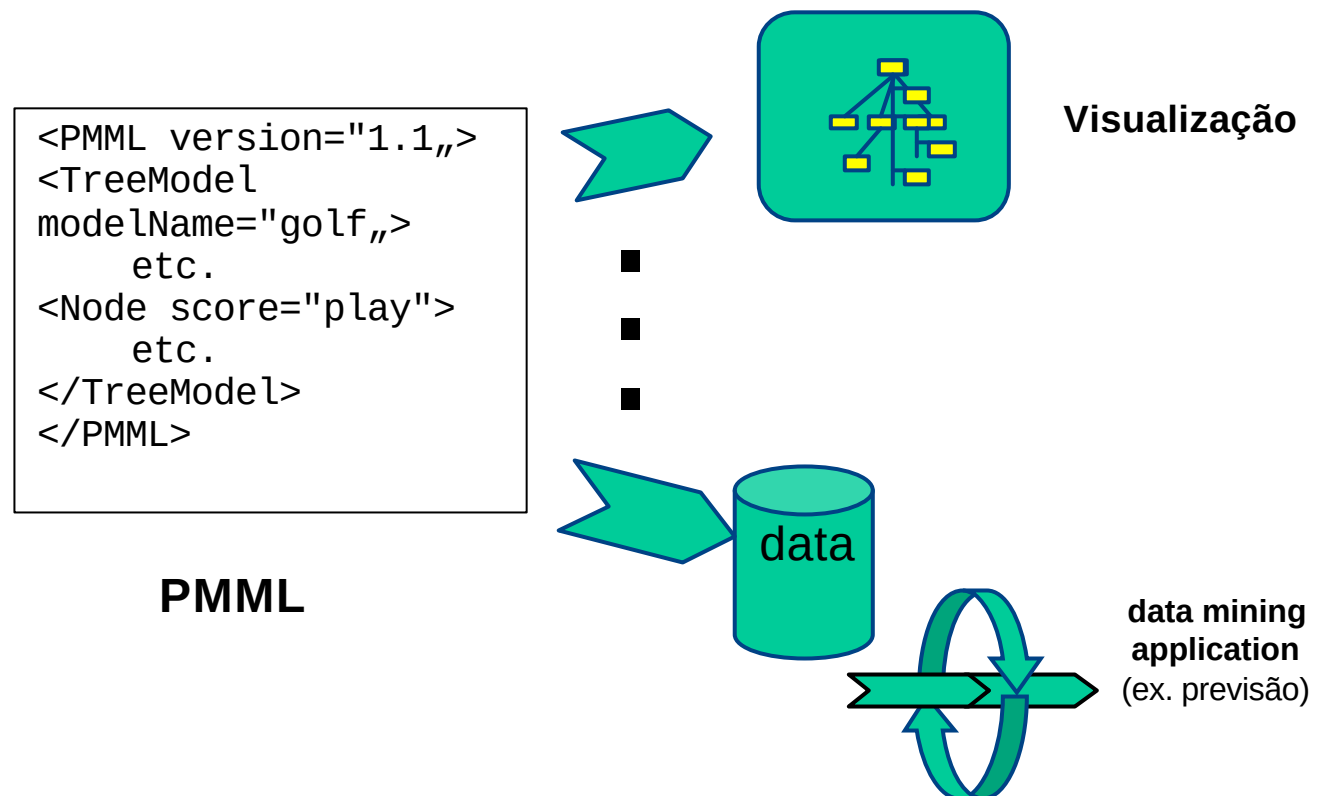
PMML (*Predictive Model Markup Language*)

Predictive Model Mark-up Language (PMML) é uma linguagem baseada no XML que permite às empresa uma forma rápida para definir modelos de previsão e partilhar esse modelos entre diferentes aplicações sem que questões de propriedade sejam levantadas.

Permite aos utilizadores criar os modelos numa aplicação, e depois visualizar, analisar e avaliar noutra aplicação.



O modelo PMML pode ser usado em diferentes contextos



Vantagens

- É um *standard* para Modelos de *Data Mining*.
- Os modelo PMML são independentes das aplicações, da plataforma e do sistema operativo.
- Não se preocupa com o processo de criação do modelo.

RDF (Resource Description Framework)

Enquanto a linguagem XML define uma estrutura, RDF permite expressar o significado associado aos dados. O significado é expresso por conjuntos de triplos que podem ser escritos utilizando marcas XML. O primeiro elemento do triplo é o sujeito, o segundo o predicado e o terceiro é o objecto. O sujeito e o objecto são identificados por URI (*Universal Resource Identifier*). O predicado é também identificado através de um URI permitindo a definição de um novo conceito ou verbo simplesmente pela definição de um apontador URI para o conceito. É muito importante a utilização de URI's diferentes para conceitos diferentes ou para variações do mesmo conceito, por forma a garantir a consistência das inferências realizadas. Por exemplo, o conceito *endereço* utilizado na linguagem corrente correspondem, na realidade, vários conceitos: pode ser um endereço de e-mail ou de um endereço físico. É, pois, necessário que sejam definidos de formas distintas e que sejam identificados por diferentes URI's.

A unicidade da relação conceito-URI garante a unicidade dos triplos de significado. Desta forma, cada conceito tem uma identificação única e os triplos constituirão uma rede de informação relacionada na qual podem usar-se regras de inferência para fazer deduções a partir do significado dos dados definido pelo triplo.[Afonso,2001]

4 Extracção de Conhecimento na Web



“São descritos todos os passos necessários e as dificuldades em extrair o conhecimento da Web. Bem como as medidas necessárias para obter resultados fiáveis numa análise.”

Extracção de Conhecimento na Web

O processo para a extracção de conhecimento dos dados da *Web* começa muito antes da utilização de uma ferramenta de *Data mining*, e continua depois da análise. Diversos passos estão envolvidos nas fases antes da análise que permitem assegurar que o processo de *Data Mining* é um sucesso e um esforço lucrativo. Apesar das ferramentas de *Data Mining* simplificarem e automatizarem a análise das bases de dados, convém serem utilizadas com cuidado para que os resultados não sejam enganosos e as conclusões erradas. Sendo assim o processo de extracção de conhecimento pode ser dividido nas seguintes etapas:

- Especificação do Problema.
- Selecção dos dados.
- Preparação dos dados.
- Avaliação dos dados
- Formatação da solução
- Selecção de ferramentas
- Construção do modelo
- Validação de resultados
- Por os resultados em prática.

4.1 Especificação do Problema

Extrair o conhecimento de um *website* envolve um planeamento avançado sobre o tipo e o nível de informação que se pretende capturar no servidor e que dados adicionais serão necessários. Isto por si só garante os esforços de *Data Mining* de forma a produzir resultados de negócio mesuráveis. Por exemplo, é necessário planear que tipo de informação (ficheiros de registo de acessos, *cookies* e formulários) se pretende capturar dos visitantes. De seguida é necessário envolver todos os campos de negócio de forma a decidir que tipo de informação é necessária para juntar aos dados vindos do servidor.

Uma planificação avançada pode poupar tempo e dinheiro na análise de *data mining*, e pode traduzir-se em resultados rápidos e assegurar o sucesso de análise de *Data Mining* do *website*. Não esquecer o princípio que ter muitos dados é melhor que não ter nenhuns. Também não esquecer quais são os objectivos de negócio e de *marketing* e qual o tipo de dados que são necessários para obtê-los.

Neste passo de planeamento são necessárias ferramentas de modelação e técnicas para construção de planos para a criação de um mapa para a etapa de

data mining. As técnicas de modelação formam uma fundação para a aplicação das técnicas de *Data Mining* e asseguram que esta seja compatível com os requerimentos dos negócios actuais. Esta modelação permite alertar para possíveis obstáculos, como a falta de certos dados necessários à análise. Outro benefício é a documentação de todo o processo de extração de conhecimento.

As seguintes ferramentas de modelação podem assegurar o sucesso do processo:

Produto	Vendedor	Website
Corporate Modeller	Casewise Systems	www.Casewise.com
ProCarta	Domain Knowledge	www.Domainknowledgeinc.com
Aris Toolset	IDS Scheer	www.Ids-scheer.com
LiveModel	IntelliCorp	www.Intellicorp.com
Workflow Modeler	Meta Software	www.Metasoftware.com
Designer/2000	Oracle	www.oracle.com
Aion	Platinum Technology	www.Platinum.com
Provision Workbench	Proforma	www.Proformacorp.com
Visio	Visio	www.Visio.com

Tabela I: Ferramentas de modelação

Identificar o potencial dos novos clientes do Website?

É provavelmente o objectivo mais comum do processo de extração de conhecimento. A classificação mais comum envolve a descoberta de atributos, características, ou perfis dos clientes do *website*. A classificação típica e a distinção dos clientes lucrativos de clientes não lucrativos.

Especificar as tendências de venda dos produtos no Website?

A descoberta das tendências ou relações entre certas páginas que representam produtos individuais ou serviços, ou seja o objectivo das técnicas de *Data Mining* é a descoberta de associações *on-line*. Este tipo de descoberta no posicionamento na *Web* de páginas, ofertas, incentivos e *links*. A descoberta de associações entre produtos únicos e serviços pode ter um impacto profundo no design do *Website*.

Identificar padrões específicos de compras no Website ao longo do tempo?

A sequenciação envolve a evolução das tendências e populações. No caso de um *website* isto representa as vendas semanais ou mensais de certos produtos ou serviços.

Para os *websites* mais visitados, como os motores de pesquisa, podem representar tendências ou padrões horários. Muitos dos problemas das previsões de *Web marketing* e de comércio electrónico envolvem a sequenciação ou a análise de séries de temporais com dados amostrados em intervalos fixos de tempo.

4.2 Selecção dos dados

Com o objectivo de negócio definido o próximo passo é escolher o servidor *Web* e os dados da empresa que satisfaz esse objectivo:

- ⚡ São os dados adequados para descrever o fenómeno de análise de *Data Mining* que pretende modelar?
- ⚡ Existe algum campo comum nos dados da *Web* que está a ser usado para ligar a outra base de dados?
- ⚡ Os dados provenientes da *Web* estão a ser consolidados com os dados do armazém de dados?
- ⚡ Se os dados utilizados na análise de *Data Mining* vão ser os mesmo e disponíveis depois da análise?
- ⚡ Que informação interna e externa está disponível para análise?
- ⚡ Os dados são actualizados e relevantes para o objectivo de negócio?
- ⚡ Os dados de padrão de vida e demográficos estão disponíveis?

Dados do Web server log

De acordo com as especificações do HTTP do W3C um *log* comum contém

- ⚡ O endereço acesso de IP do cliente
- ⚡ Data e hora de acesso
- ⚡ URL visitado
- ⚡ Modo de pedido (GET, POST, HEAD)
- ⚡ *browser* do visitante
- ⚡ *login* e *password*, se o servidor requer autenticação

```
128.101.228.20 - - [10/Nov/1999:10:16:39 -0600] "GET / HTTP/1.0" 200 - "-"  
"Mozilla/4.61 [en] (WinNT; I)"  
128.101.228.20 - - [10/Nov/1999:10:16:39 -0600] "GET /HEAD-BKGRND.jpg  
HTTP/1.0" 304 - "http://www.cs.umn.edu/" "Mozilla/4.61 [en] (WinNT; I)"  
128.101.228.20 - - [10/Nov/1999:10:16:39 -0600] "GET /HEAD-UMN-LOGO.gif  
HTTP/1.0" 304 - "http://www.cs.umn.edu/" "Mozilla/4.61 [en] (WinNT; I)"  
128.101.228.20 - - [10/Nov/1999:10:16:39 -0600] "GET /head-bar.jpg HTTP/1.0"  
304 - "http://www.cs.umn.edu/" "Mozilla/4.61 [en] (WinNT; I)"  
128.101.228.20 - - [10/Nov/1999:10:16:39 -0600] "GET /cs-building.gif HTTP/1.0"  
304 - "http://www.cs.umn.edu/" "Mozilla/4.61 [en] (WinNT; I)"  
128.101.228.20 - - [10/Nov/1999:10:16:39 -0600] "GET /new.gif HTTP/1.0" 304 -  
"http://www.cs.umn.edu/" "Mozilla/4.61 [en] (WinNT; I)"  
128.101.228.20 - - [10/Nov/1999:10:16:39 -0600] "GET /tail-bar.jpg HTTP/1.0" 304 -  
"http://www.cs.umn.edu/" "Mozilla/4.61 [en] (WinNT; I)"
```

FIGURA 16: EXTRACTO DE DE UM FICHEIRO DE REGISTO DE ACESSOS

Dados de descrição do site

Os dados de descrição do *site* permitem descrever não só a estrutura de *links* existem, mas também a informação contida no próprio *site*.

- Descrição da topologia do *site* através da estrutura de *links* do *site*

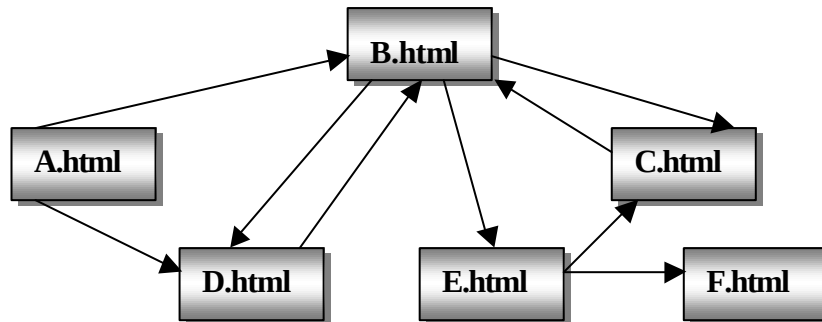


FIGURA 17: DESCRIÇÃO DO SITE [Spiliopoulou, 99]

- Os metadados em cada página *Web*, como:

- ✎ registos de descrição da página;
- ✎ *Tags* de HTML da página;
- ✎ *Tags* de XML da página.

Dados Hierárquicos

As hierarquias reflectem e generalizam as semânticas dos objectos podem ser construídas baseadas no:

- ✎ Conteúdo das páginas *Web* – a hierarquia é criada com a informação contida nas *tags* da página, na relação de produtos das páginas;

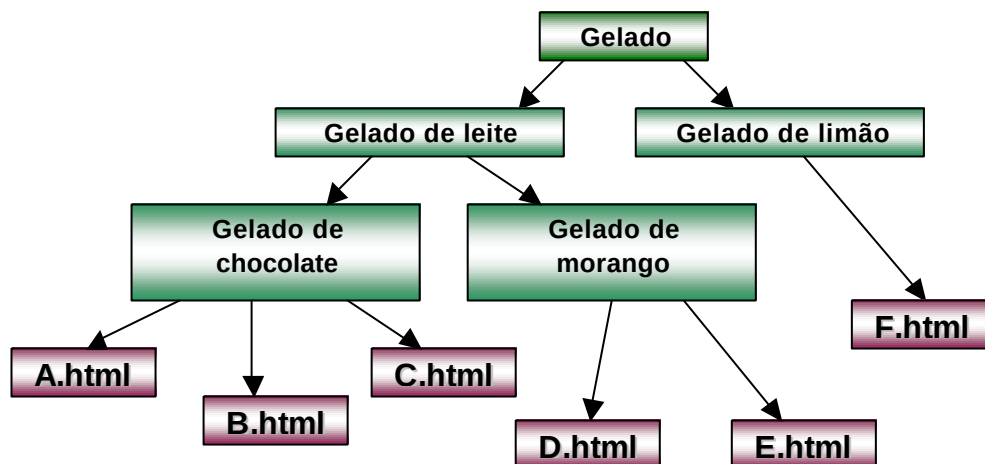


FIGURA 18: HIERARQUIA DO CONTEÚDO DAS PÁGINAS [Spiliopoulou, 99]

➤ Nos *hosts* dos utilizadores, este podem ser organizados hierarquicamente;

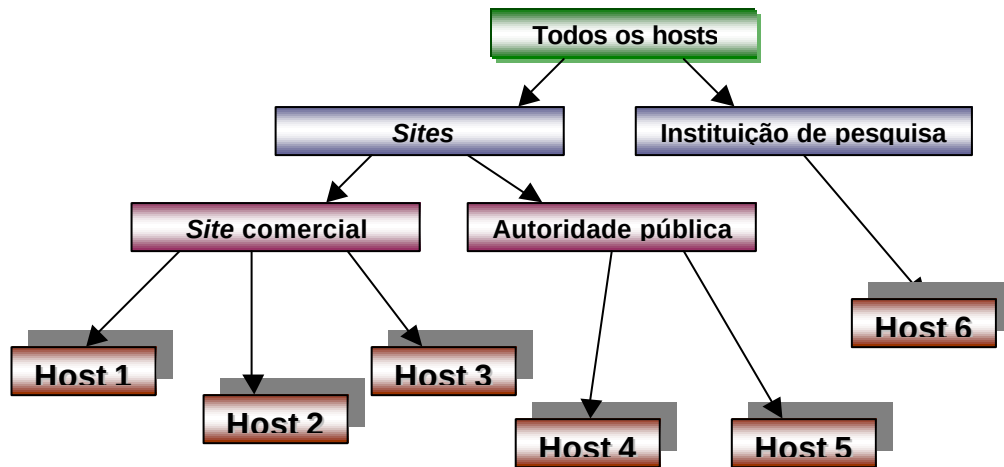


FIGURA 19: HIERARQUIA DE UTILIZADORES [Spiliopoulou, 99]

➤ No tipo da página de Web;

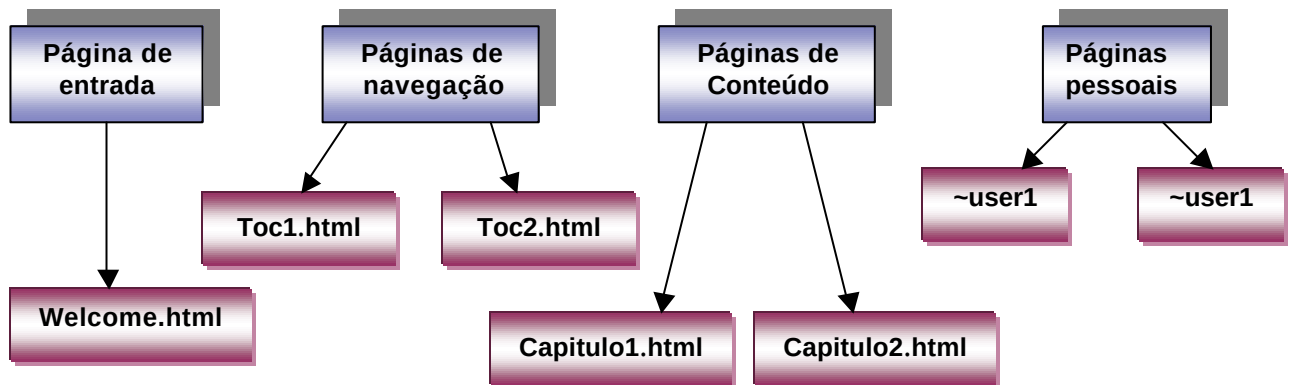


FIGURA 20: HIERARQUIA DE PÁGINAS [Spiliopoulou, 99]

Dados das transacções

Quando se constrói um armazém de dados, há que ter em mente que os dados provêm muitas vezes de diversos sistemas e que muitas vezes isto envolve lidar com *flat-files*, tabelas relacionais, e sistemas hierárquicos e relacionais de múltiplos sistemas operativos que lidam ao mesmo tempo com múltiplos atributos (como as datas) de maneiras diferentes. A coordenação e selecção destes conjuntos de dados envolve considerável esforço e tempo sendo esta parte que demora mais tempo de todo o processo de extracção de conhecimento.

Os factores decisivos nas decisões da selecção dos dados da *Web* são a sua portabilidade e o seu custo quer das bases de dados internas como as externas. Outros factores que intervêm na selecção são a segurança e a privacidade dos dados, especialmente no uso da informação cedida pelos

clientes do *website*. Não esquecer que alguma informação tem que ser utilizada para a criação do perfil do utilizador, e esta pode não estar disponível.

Ao longo de todo este processo aparecem dados inconsistentes, e diversos tipo de dados encriptados, e valores inconsistentes quando seleccionados e comparados com dados de outras fontes. Isto pode prejudicar todo o processo de extracção de conhecimento levando a soluções erradas.

A necessidade de certos dados para a análise de *Data Mining* estão directamente ligados com o *output* ou *insight* desejado. Para a maior parte das análises, a regra geral é “quanto mais dados melhor”. A segunda regra é que uma selecção de dados da *Web* deve conter ambos os tipos de clientes ou contas, ou seja tanto os casos positivos como os casos negativos, para que a previsão seja a mais realista possível.

4.3 Preparação dos dados

Uma vez que os dados foram reunidos e visualmente inspeccionados, algumas decisões tem que ser tomadas em relação aos atributos a serem excluídos e aos atributos a serem convertidos em formatos utilizáveis.

- Qual é a condição dos dados, e quais os passos necessários para prepará-los para análise.
- Quais as conversões e mapeamentos de dados necessários antes da análise.
- Só são processos de preparação aceitáveis para os utilizadores se produzirem soluções.
- Quais as estratégias que se pretendem tomar para lidar com os dados que faltam e com os dados isolados.
- Se são necessárias conversões para os campo categóricos.
- Normalização de campo monetários.
- Se são necessárias conversões de datas para valores contínuos.
- Conversões de moradas.
- Conversões de campos Sim/Não para 1/0.

Uma ferramenta gráfica ou um bom editor de texto podem assistir na inspecção física dos dados. Uma inspecção visual permite uma visão geral do número e da percentagem de campos em branco no conjuntos de dados. Também uma ferramenta de estatística pode assistir na identificação de relações importantes entre as variáveis existentes nos dados. No entanto, isto pode não ajudar em conjuntos de dados muito grandes. Quando se utiliza bases de dados muito

grandes, deve-se extrair uma amostra aleatória, e executar as operações, extrair conhecimento dos dados e avaliar os resultados.

Se consideramos a utilização de ferramenta de redes neuronais, para além das conversões das datas, têm que ser feitas conversões dos valores categóricos para valores de 1 até N e todos os valores são convertidos para valores compreendidos no intervalo de 0 a 1 ou em funções logarítmicas ou de raiz quadrada. Um dos métodos mais comuns para tornar os dados mais homogéneos, é a redução do números de valores distintos para uma dada característica.

O dados que faltam, apresentam um problema, especialmente quando se utilizam redes neuronais, porque esta não lidam com valores em branco. Os dados que faltam têm que ser analisados para saber se eles representam uma resposta negativa que precisa de ser corrigida. Pode-se substituir os campo em branco por zero, mas este valor não representa a realidade, o que produz um modelo pouco realista.

4.3.1 Melhorar os dados

Uma das estratégias utilizadas na preparação dos dados para análise de *Data Mining* é a ligação a outras bases de dados existentes para a produção de ficheiros:

Tipos de ficheiros	Possível conteúdo
Dados da transacções	Tipos de serviços ou produtos vendidos
Códigos postais e disposição demográfica dos clientes	Sexo, Idade, Nível de renda, Filhos
Outbound Mail	Taxa de resposta, Número de Mailings
Informação do consumidor	Número total de vendas, é cliente desde (Data)
Armazéns de dados	Perfil do cliente por produto e serviço
Datamart	Perfil do cliente por divisão ou região
Relatórios de credito do consumidor	Cartões de Credito/Debito, ocupação, hipoteca

Tabela II: Tipos de ficheiros

4.3.2 Remoção de entradas indesejáveis

O acesso relevante a uma página depende do objectivo da análise:

Optimização do desempenho do servidor

- Todas as entradas são relevantes.

Estudo da utilização do site

- O acesso às imagens devem ser removidas.

Estudo do comportamento dos utilizadores

- O acesso às imagens devem ser removidas;

- Acesso por parte de robôs¹ deve ser removido.

4.3.3 Distinção entre os utilizadores

As actividades intervaladas de múltiplos utilizadores do mesmo computador (ou *proxy*) são guardadas como actividades de um único utilizador.

4.4 Avaliação dos dados

Uma avaliação estrutural dos dados é feita para a determinar as ferramentas de *data mining* a serem utilizadas para análise.

- Qual o rácio de atributos categóricos/ binários na base de dados?
- Qual a natureza e estrutura da base de dados?
- Qual é a condição geral do conjunto de dados?
- Qual a distribuição do conjunto de dados?
- Os dados do conjunto de dados são actuais?

Como regra geral, as redes neuronais funcionam melhor em conjuntos de dados com um grande número de atributos numéricos. O algoritmos de *Aprendizagem automática* incorporados nas árvores de decisão e as ferramentas de *data mining* geradoras de regras funcionam melhor com conjuntos de dados que tenham um grande número de registos e um largo número de atributos. A escolha de um tipo de algoritmos tem um impacto profundo na exactidão dos resultados do *data mining*.

4.5 Formatação da Solução

Junto com a avaliação dos dados da *Web* e os objectivos de negócio aparece a questão da forma a dar à solução.

Existem diversas formas que a solução pode tomar, desde árvores de decisão, regras em código C, gráficos, mapas. Tendo sempre em conta o objectivo da solução (classificação, regressão, *clustering*, segmentação). Depois do processo de extracção de conhecimento é necessário definir como é que a informação será distribuída e em que formatos estará disponível.

Múltiplas ferramentas serão necessárias para chegar a um formato ideal para um *website*. Por exemplo quando é necessário extrair regras de uma análise de *clustering*. Para isso primeiro é necessário um formulário da análise de

¹ O acesso por robôs pode ser reconhecido pelo nome e pelo seu comportamento estranho: acesso repetido a mesma página, acesso consecutivo em intervalos muito curtos de tempo estes robôs são utilizados para efectuar transacções na web ou para pesquisar informação.[Spiliopoulou, 99]

clustering, usando *Self-Organization Map* ou as redes de *Kohonen*. De seguida correr os *clusters* encontrados por um algoritmo de *Aprendizagem automática* para gerar regras IF/THEN dos *clusters*. Depois analisar com um algoritmo de aprendizagem automática num conjunto de dados com grande número de atributos para que estes sejam reduzidos, ou seja para chegar a um grupo de atributos que sejam significativos para análise, e depois correr os atributos encontrados por um algoritmo de rede neuronais para determinar o modelo de classificação final. Por isso é necessário pesar muito bem quais as opções do negócio e do *website* e quais as ferramentas necessárias à análise, e o tipo de análise que se procura.

Muitas vezes o formato da solução de *data mining* determina qual a ferramenta de *data mining* a utilizar:

- Se é necessária uma explicação como os padrões nos dados da *Web* é necessário um algoritmo de aprendizagem automática como as árvores de decisão ou um gerador de regras
- Se for necessário precisão e eficiência é mais adequado um algoritmo de redes neuronais.

Para um *site* de *e-commerce* ambos os casos descrito acima serão vantajosos. Neste caso, conhecer a demografia dos clientes e o aumento das vendas são os objectivos desejados.

4.6 Selecção de Ferramentas

A selecção das ferramentas de *data mining* dependem muito do objectivo que se pretende atingir. Por exemplo, a seguinte tabela relaciona os tipos de ferramentas de *data mining* com o resultado pretendido.

Tipos de Ferramentas	Classificação	Clustering	Descrição
Algoritmos Genéticos	Sim	Não	Não
Aprendizagem automática	Sim	Sim	Sim
Redes Neuronais	Sim	Sim	Não
Self-Organizing Maps	Não	Sim	Não

Tabela III :Tipos de Ferramentas

Quando se escolhe uma ferramenta de *data mining* também tem que se ter em conta os seguintes factores:

- Número de campos com valores contínuos;
- Número de variáveis dependentes;
- Número de campos categóricos;
- O comprimento e tipo de registos.

Em regra, os algoritmos de aprendizagem automática têm melhor desempenho em conjuntos de dados com um grande número de atributos categóricos e com grande número de campos por registo. Os algoritmos de redes neuronais funcionam melhor em campos numéricos.

4.6.1 Escalabilidade

À medida que os ficheiros de registos de acessos e os registos das base de dados aumentam, também o desempenho da ferramenta utilizada deve aumentar. Escalabilidade significa que tirando partido das base de dados paralelas de sistema e dos processadores adicionais, o utilizador é capaz de trabalhar com mais dados, construir mais modelos, e melhorar a precisão geral adicionando processadores. Com o aumento dos dados também aumenta a complexidade, não só no número de registos mas também no número de atributos, variáveis, e padrões do *website*. Se a plataforma utilizada não tem a possibilidade de escalar, a natureza computacional da ferramenta de *data mining* vai atrasar e eventualmente matar o sistema de suporte à decisão.

4.6.2 Precisão

A precisão é medida pela taxa de erro do algoritmo que prevê os modelos. Existem diversas formas de avaliar a previsão. Pode ser medida como o grau de erro, que determina até que ponto que as respostas foram erradas ou se houve alguma resposta. Na selecção de uma ferramenta deve-se procurar saber:

- Como é que o algoritmo da ferramenta determina se as características encontradas se enquadram ao problema?
- Que providências são tomadas para tratar os dados em falta, ruído e os custos computacionais?
- Se trata as variáveis dependentes e independentes?
- Qual a quantidade de ruído suporta até baixar a precisão?
- Qual a sensibilidade do algoritmo ao ruído e aos dados que faltam?

4.6.3 Formatos

Uma ferramenta de *data mining* que incorpora um algoritmo de *aprendizagem automática* deve permitir a visualização das regras mais importantes ou dos factores que descobriu. A ferramenta deve não só detectar os padrões, mas também dar uma ideia do que se está a passar no *website* e permitir uma interacção fácil com os dados. A explicação da actividade do *website* consiste em saber qual a natureza das transacções dos clientes que estão a ser analisados. A ferramenta de *data mining* deve gerar os resultados em árvores, fórmulas, gráficos ou regras, devendo estes serem o mais simples possíveis, para que se possa descobrir o maior número de objectos escondidos. Para um conjunto de dados de tamanho considerável, como um de um *website* de vendas, a descoberta de todas as regras seria impossível para um utilizador, mesmo que todas as regras fossem simples de compreender. As regras gerais devem descrever qual a actividade que está a acontecer no *website* a alto nível.

Se a ferramenta incorporar prioridades de resultados deve ter-se em atenção como é utilizado, qual o seu nível importância, qual a percentagem, probabilidade e precisão de êxitos, e qual o número de casos que incorporam nas regras. A ferramenta pode também oferecer formas de visualização, como histogramas de erros, ou localizar em mapas geográficos segmentos de mercado. A ferramenta pode ser desenhada para trabalhar no *desktop* do utilizador ou em ambiente cliente/servidor.

4.6.4 Soluções

As ferramentas de *data mining* devem ajudar a compreender os resultados da análise dando medidas simples, mas exactas, como de previsão da taxa de erros, nível de significância e/ou a taxa de precisão.

Os tipos de *output* que a ferramenta gera são: árvores de decisão, código C, regras IF/THEN, relatórios narrativos, gráficos, sintaxe SQL. Algumas ferramentas produzem resultados que são relativamente simples de interpretar, como as árvores de decisão que podem ser traduzidas em regras. Outras ferramentas, como redes neuronais, são muito boas nas previsões, mas são difíceis de compreender. Sendo assim o formato dos resultados de uma ferramenta devem ser balanceadas entre a compreensão e a precisão.

Uma ferramenta deve também incorporar instrumentos adequados que permitam ao utilizador executar análises sensíveis aos resultados. A ferramenta também podem incluir algum interface de exportação dos resultados dos modelos para outras ferramentas de visualização ou ferramentas de OLAP.

4.6.5 Pré-processamento

A preparação dos dados para análise é uma das etapas que consome mais tempo em todo o processo de *data mining*. Geralmente, os ficheiros de registo de acessos são delimitados por vírgulas. Sendo assim, a ferramenta utilizada deverá poder utilizar este formato sem dificuldade. Os resultados dos formulários do *website* devem ser, provavelmente, armazenados em bases de dados relacionais, que a ferramenta utilizada deve ler sem problemas.

Qualquer que seja a característica que as ferramentas tenham, estas devem tornar mais simples e fácil todo o processo de extracção de conhecimento e de construção de modelos de previsão.

Algumas funções a considerar na avaliação das ferramentas a utilizar:

- limpeza dos dados;
- selecção dos dados;
- descrição dos dados;
- transformação dos dados.

A característica chave de uma ferramenta de *data mining* é a sua capacidade de lidar com os dados autonomamente, ou seja, como eles são apresentados, executando algumas filtragens, como por exemplo os tratamento dos dados que faltam, oferecendo hipóteses de descartar, fazer médias, alertar, e excluir, etc.

A ferramenta deve ser capaz de dar um relatório descritivo da qualidade dos dados utilizados. Também há que ter em conta os seguintes factores:

- Como é que os dados são descritos, contagem de linhas e de valores, a distribuição de valores.
- Se consegue distinguir a diferença entre valores contínuos e campos categóricos.
- Se a ferramenta reconhece e trata os dados de acordo com o seu tipo (ex. Inteiros, Reais)
- Como é que a ferramenta trata os dados que faltam e se identificam violações de integridade.
- Se a ferramenta consegue executar alterações directamente nos dados.
- Se consegue agrupar variáveis em amplitudes.
- Se a ferramenta executa cálculos nas colunas existentes.
- Se a ferramenta requer extensiva preparação de dados.

4.6.6 Conexões

Algumas ferramentas de *data mining* requerem extracção dos dados a serem analisados, das bases de dados para o formato interno de ficheiros, enquanto outras ferramentas trabalham directamente com as bases de dados.

Uma ferramenta de *data mining* que acede directamente ao servidor Web utilizando SQL maximiza o desempenho e tira partido do acesso a bases de dados paralelas.

4.6.7 Import/Export

As ferramentas de *data mining* oferecem geralmente integração com outros produtos e têm formas de obter a totalidade dos dados analisados no processo de descoberta de conhecimento. Relacionando a importação dos dados com a conexão e se a ferramenta suporta múltiplos formatos, como ASCII, MDB, XSL, ficheiros delimitados por virgula ou *tab*, SAS, SPSS.

É preciso ter em atenção quais são as conversões que a ferramenta faz com os dados e qual a taxa de importação e a ferramenta permite a exportação do código, sintaxe, e regras. A ferramenta deve ter a capacidade de facilmente ligar os resultados a um formato que possa ser exportado.

4.6.8 Administração de memória

Normalmente a memória que uma ferramenta de *data mining* requer depende do tamanho do conjunto de dados. O tempo para completar a fase de descoberta de conhecimento é um factor limitador para a maior parte das ferramentas de *data mining*. Se a utilização de um algoritmo o seu tempo de execução cresce exponencialmente no tempo, o tamanho do conjunto de dados de treino será bastante limitado.

Quando se considera a utilização de memória por parte de uma ferramenta de *data mining* só a complexidade interessa. Sendo assim, a utilização da memória pode dar uma indicação do sistema necessário para tratar quantidade “normais” de dados, em termos de registos.

4.6.9 Desempenho

A velocidade e a precisão contribuem para a avaliação do desempenho de uma ferramenta de *data mining*. A velocidade é medida pela rapidez com que o modelo é construído, e a rapidez com que o modelo de previsão pode avaliar novos dados. Outro factor que tem um impacto no desempenho é o custo. Qual é o custo de arranjar um conjunto de dados e do desenvolvimento do modelo de previsão? Isto inclui o custo do número dos exemplos necessários e o custo de assegurar a precisão de um conjunto de dados de um modelo. Por exemplo, no volume de transições de um *website* este problema não se apresenta.

4.6.10 Ruído

A precisão da ferramenta é varias vezes afectada pelo ruído, que é o resultado de colunas irrelevantes e de falta de dados. Na avaliação da robustez de um produto de *data mining*, deve ter-se em conta a quantidade de ruído que a ferramenta pode lidar até que a sua precisão baixe. Em vez de avaliar a ferramenta de *Data Mining* com um conjunto de dados de treino perfeito, deve ser feito com um conjunto de dados que tenha em conta todo o ruído do sistema.

Diferentes níveis de ruído podem ser adicionados ao conjunto de dados de treino e registar-se o número de classificações correctas encontradas. Ao adicionar o ruído deve-se estudar as inconsistências que resultam da análise.

4.6.11 Paradigmas

Um dos pontos chave para a selecção de uma ferramenta de *Data mining* são as tecnologias de *data mining* que suporta como por exemplo ser capaz de executar as rede de *Kohonen* para descobrir os *clusters* e gerar regras usando um algoritmo de *aprendizagem automática* (C5.0), tudo isto dentro de uma ferramenta de *data mining*. Na escolha de uma ferramenta deve ter-se em conta o tipo de tecnologias que suportam para que os resultados obtidos sejam optimizados. Sendo assim, na escolha de uma ferramenta também se deve ter em consideração os seguintes factores:

- Qual a tecnologia ou tecnologias que suportam;
- Se a ferramenta consegue dividir os dados em conjuntos de treino e de teste e se o utilizador consegue introduzir a percentagem de divisão que quer;
- Se a análise é automatizada ou se permite algum nível de personalização;
- Se a ferramenta de *data mining* incorpora mais do que um algoritmo;
- Se a interface permite uma avaliação dos resultados da análise;
- Se a ferramenta é de fácil utilização.

4.6.12 Eficiência

O ultimo critério para a selecção de uma ferramenta é a eficiência. Por exemplo quanto tempo é que demora o processo de extracção de conhecimento a chegar a uma conclusão, a facilidade de chegar rapidamente a um regra geral.

4.7 **Construção do Modelo**

É só nesta etapa que a extracção de conhecimento do *website* começa. Este é o processo que procura os padrões num conjunto de dados e gera a classificação de regras, árvores decisão, *clustering*, a determinação de resultados, de pesos e a avaliação e comparação de taxas de erro.

O desenvolvimento de um modelo deve permite analisar tanto os casos negativos como os casos positivos, de um conjunto de dados, por exemplo os clientes compradores dos não compradores.

Em alguns casos, especialmente nos grandes *sites* de venda, existem milhares de amostras a partir dos quais se podem desenvolver modelos que ajudam a prever situações futuras. No entanto, num *site* de vendas existem múltiplos produtos ou serviços, o que trona necessária a construção de modelos separados para cada um.

Se não existem amostras negativas ou positivas, o método mais utilizado nestes casos é o *clustering* ou a aprendizagem não supervisionada. Este método torna-se mais difícil que a classificação, mas torna-se absolutamente necessário. O *clustering* pode ser utilizado para a descoberta de padrões de consumidores. Os revendedores também utilizam muitas vezes regras de associação para definir a posição de certos produtos que tende a ser vendidos em conjunto. Por exemplo, os consumidores que compram vinhos também compram queijos. Padrões semelhantes também podem ser descobertos num *website* comercial. Por exemplo, os produtos que são comprados ao mesmo tempo e as sequências de compra, ou seja, um consumidor compra um produto hoje e dias mais tarde volta comprá-lo.

Existem hoje em dia, procuras exaustivas de associações de compra e de padrões e venda nos grandes *sites* de vendas (livros, CDs, brinquedos e *software*). As associações descobertas levam a oportunidades de vendas conjuntas e promoções.

As aprendizagens supervisionadas e não supervisionadas requerem um conjunto de dados de treino e teste. A divisão dos dados originais pode ser feita utilizando um número aleatório. Quase todas as ferramentas modernas de *data mining* incorporam métodos para divisão dos dados e avaliação dos resultados depois do modelo criado.

Uma vez o modelo criado a partir do conjunto de dados de treino, os dados de teste são passados pelo modelo para avaliar a precisão do mesmo. Dividindo o número de classificações incorrectas pelo número de classificações totais obtém-se a taxa de erro. Dividindo o número de classificações correctas pelo número de classificações totais obtém-se a taxa de precisão do modelo (que será 1-taxa de erro).

Quase todas as ferramentas actuais permitem medir a precisão do modelo construído. Elas facilitam o processo da divisão dos dados, em dados de treino e teste. Também permitem a comparação e teste aleatório dos resultados encontrados dos dados de treino com os dados de teste. É muito importante que os dados de teste sejam dados que não entraram para a construção do modelo para que os resultados sejam mais credíveis e por isso se dá a separação dos dados originais.

4.8 Validação dos resultados

A análise de um *website* envolve diversos departamentos, como sistemas de informação, *marketing*, vendas, inventário, etc, e terá que ter um conjunto de pessoas como administradores, *designers*, analistas, e os responsáveis pela manutenção diária das operações do *website*. É importante que todo este conjunto de pessoas esteja a par dos resultados encontrados na análise de *data mining*. Também é conveniente que especialistas na área analisem os resultados para que assegurem que estes estejam correctos e apropriados aos objectivo de negócio do *site*. Se os resultados da análise foram falsos, seja por causa dos dados utilizados, seja por causa da ferramenta ou da metodologia, será necessário outra análise e a construção de novos modelos

A análise dos resultados também permitem responder a perguntas como:

- Se importantes relações forem encontradas entres as variáveis dependentes e independentes. Exemplo: se as regiões demográficas descobertas são as mesmas das análises anteriores.
- Se o *site* teve outros clientes além dos esperados.
- Se consoante o tempo o número de pessoas que estiveram “ligadas” aumentaram.

➤ Se por causa de uma promoção o número de vendas aumentou.

4.9 Entrega dos Resultados

O relatório de *Data mining* deve ser preparado, documentando todo o processo de descoberta de conhecimento, incluindo todos os passos tomados na selecção e preparação dos dados, ferramentas utilizadas e porquê, as configurações efectuadas, os resultados e explicação do código gerado.

Como em qualquer área de negocio é necessário estabelecer regras gerais e procedimentos. No relatório terá que haver comentários como foram obtidos os resultados de *data mining*, a começar por verificar se o objectivo do *website* foi conseguido ou não, e em caso afirmativo deve-se explicar o porquê. Também deve ser incluído no relatório como é que os resultados de *data mining* podem ser melhorados. Pode-se mencionar as ajudas que auxiliaram o processo e quais os dados adicionais que podem auxiliar a análise, recomendações para a captura de informação como, por exemplo, melhoramentos aos formulários e as *cookies* de registo do *website*. Devem ser incluídos algumas recomendações para mudanças subtis ao *design* do *website*.

O processo de descoberta de conhecimento é um processo interactivo, e a sua documentação deve assegurar que os melhoramentos tenham um impacto positivo, especialmente se o *site* for de comércio electrónico.

4.10 Por os resultados em prática

Esta é a etapa final do processo de extracção de conhecimento dos dados de um *website*. Esta etapa envolve a incorporar os resultados encontrados na análise, nos objectivos de negócio, nas campanhas de *marketing* e no planeamento estratégico de empresa.

4.10.1 Integração do website e Armazéns de dados

A integração dos dados do *website* com a informação da empresa ou com a informação de um armazém de dados pode ajudar na actualização constante do mesmo, de forma a manter a satisfação dos clientes actuais e de forma a conquistar novos clientes.

Com a integração do *website* com os resultados das análises as bases de dados das empresa, os clientes têm acesso a informação muito mais actualizada sobre inventários, serviços e suporte a clientes. Consoante as visitas dos clientes aprende-se as suas preferências sobre produtos e serviços. Isto é feito a partir da extracção de conhecimento dos dados reunidos sobre os visitantes ao *website*, e este pode começar a disponibilizar informações específicas a determinado tipo de visitantes consoante o perfil em que eles se inserem.

As empresas podem assim comunicar com os clientes de um forma muito personalizada. E com a integração com os dados das compras efectuadas o cliente pode ser informado de produtos similares, e fazendo um cruzamento de vendas relacionadas pode dar-se a informação de serviços complementares.

Existem diversos factores que podem ser levados em consideração durante este processo como as datas das compras e o valor total das vendas efectuadas no *website*, estes factores podem servir para classificar o cliente consoante o seu valor/lucro, e a sua lealdade.

Com a integração do *website* com outros tipos de dados, pode ser possível a identificação dos clientes mais lucrativos e a identificação de potenciais clientes. Com base nos perfis efectuados o *website* pode alvejar novos clientes enquanto eles visitam o *site*, com a comparação de vários atributos entre os clientes e os visitantes de um *site* de forma a adicionar e alterar ofertas, incentivos, preços, *banners* e outras comunicações. Também podem ser feitas alterações a campanhas de *marketing* e no *design* do *website*. Por exemplo os clientes podem ser comparados, a visitantes do *site* nos seguintes atributos:

- ~ Idade;
- ~ Sexo;
- ~ Rendimento;
- ~ Passatempos;
- ~ Ocupação;
- ~ Estado civil;
- ~ Residência;
- ~ Habilitações literárias;
- ~ Interesses lúdicos;
- ~ Taxa de rendimento disponível;
- ~ Posicionamento demográfico e psicográficos;

Os dados psicográficos caracterizam as atitudes de um indivíduo e caracteriza o seu estilo de vida. Estes dados são um indicador de como os clientes e os visitantes pensam do seu estilo de vidas.

O SRI Internacional, um instituto de pesquisa não lucrativo, desenvolveu oito categorias de padrão de vida para os consumidores. Assim os consumidores são agrupados em grupos mutuamente exclusivos baseados na sua psicologia e diversos factores demográficos. Olhando para o lado humano da equação, por exemplo as preferências por certos valores tais como a liberdade, tradição, novidade, informação ou simulação. Aparecem novos grupos que são úteis na avaliação da qualidade dos ambientes para determinados utilizadores. Assim podem criar-se perfis de utilizadores da *Web* e criar *clusters* partindo dessas categorias.

Esta etapa final da extracção do conhecimento requer uma monitorização dos modelos gerados. Todos os modelos gerados com o tempo iram tornar-se obsoletos sendo por isso a monitorização tão importante para manter a precisão.

A análise de *Data mining* pode sugerir algumas alterações a nível da organização e como os recursos do *website* são desenvolvidos, ou quais as recomendações que serão feitas à gerência, e quais as alianças e sociedades que serão consideradas. A integração dos resultados da análise de *data mining* no *website*, vai ter um impacto na forma como o inventário vai ser organizado *on-line*, ou como as promoções são planeadas, ou quais os visitantes ou clientes que terão determinadas ofertas. Isto também poderá envolver a integração de um sistema de e-mail para uma notificação personalizada de produtos ou serviços.

5 Ferramentas



“As ferramentas existentes no mercado para extrair conhecimento da Web. As áreas para que estão vocacionadas desde as análise dos perfis de clientes a análises de associações de produtos.”

5.1 Visão Geral

As ferramentas de Data Mining completam outros *software* para análise de dados. As ferramentas estatísticas, as ferramentas de visualização de dados, sistemas de informação geográfica, OLAP. As ferramentas de *Data Mining* fazem parte do conjunto de interfaces para extrair conhecimento dos. Cada uma com suas funcionalidades complementares, adequadas a diferentes tipos de utilizadores e de análises.

Enquanto as ferramentas tradicionais das bases de dados são capazes de mostrar o que está na base de dados, o *software* analíticos ajudam o utilizador a descobrir muito mais do "porquê?" dos acontecimentos. Ao passo que numa ferramenta estatística, o utilizador pode apenas gerar hipóteses as ferramentas de *Data Mining* estende a capacidade de gerar e validar hipóteses descobrindo conhecimento.

Resumindo *Data Mining* é um processo indutivo de análise. Quando os dados são demasiado grandes e complexos para serem analisados, a diminuição a forma de rácio ou formulas, que podem levar a padrões. O grande volume de dados gerados de um *website*, muitas vezes escondem padrões que revelam as condições quando os clientes compram ou clicar em certos anúncios ou *banners*.

O *software* de *Data mining* pode revelar como um valor de um campo numa base de dados é afectado por valores de outros campos. Por exemplo um campo como "n.º total de compras" pode ser afectado pelos valores dos outros campos do ficheiro de registos de acessos ou dos formulários das bases de dados, como a idade do cliente, sexo ou as palavras chave utilizadas. Neste exemplo "n.º total de compras" pode ser explicado previsto enquanto os outros campos serão o *input* para o modelo de *Data Mining*.

Todas as ferramentas aqui descritas tem por base as páginas das empresas que as produzem. Neste capítulo são dados exemplos de ferramentas de *data mining* que podem analisar dados da *Web* tal como o Clementine e ferramentas que são exclusivamente dedicadas ao *web mining* tal como WebAnalyst. São dados também exemplos de ferramentas que apesar de não serem ferramentas de *data mining*, permitem extrair algum conhecimento através das estatísticas que fornecem. Outras ferramentas estão descritas em anexo.

**DATA MINING
SOFTWARE**

5.2 Blue Martini Marketing



<http://www.bluemartini.com>

Blue Martini
Marketing

Blue Martini Marketing

Blue Martini Marketing é uma solução de análise para clientes e de automação de marketing. As companhias usam o *software* para criar um quadro unificado de clientes, a análise do comportamento de clientes para a padronização e previsão da personificação de site e de campanhas de *marketing*. Em particular para o incremento do lucro e do retorno dos programas de *marketing*

Segmentação do mercado para atingir os clientes alvo

Blue Martini Marketing emprega os armazéns de dados para criar um perfil de clientes. As ferramentas de análise todos os perfis dos clientes permitindo segmenta-lo e refinar a estratégia de *marketing*. Com base na análise efectuada podem ser atingidos os clientes mais valiosos e maximizar o custo efectivo de cada cliente.

Executar campanhas de custo efectivo para adquirir mais clientes

Os resultados da análise melhoram drasticamente os custo efectivo do programas de *marketing* pelo aumento a resposta a taxa conversão. A aplicação executa campanhas de *e-mail* e *direct mail* a clientes alvo pré-seleccionados pela análise

As interacções personalizadas levam a um rendimento

A análise baseada em *marketing* aumenta a retenção de clientes com interacções obrigatórias. Porque adquirir um cliente é muito mais caro que reter um cliente, o investimento é recompensado.

Cria um grupo virtual para um rápido *feedback*

As campanhas de *marketing* podem incluir controlo sobre grupos para medir o impacto e sugere políticas de ajuste aos grupos. Para um mais rápido *feedback*, campanhas podem ser usadas para criar grupos que podem fazer teste de mercado que podem testar desde mensagens de publicidade a *designes* de novos produtos.

Blue Martini Software
Module Detail**Data mining**

O modulo de *data mining* fornece uma análise das capacidades para compreender os clientes e encontrar informação relevante no meio de um amontoado de dados. *Data mining* expõem subtis padrões que prevêem e levam a personalização em todos os pontos de contacto. Fecha uma ciclo entre análise e as vendas que prevêem a produção de mecanismos que geram lucro e vantagens competitivas.

O valor das previsões

O *Data mining* invoca algoritmos que identificam regras de significados estatísticos e associações escondidas na história das relações dos clientes. *Mining* produz recomendações não obvias baseadas no cruzamento dos dados. Também gera modelos que prevejam tendência de um indivíduo para comprar baseados nos dados demográficos e no historial de encomendas. A personalização do modulo é feita por regras associações e modelos para apresentação aos indivíduos de conteúdos, ofertas, promoções, para maximizar as vendas

Resultados

Quando são analisados os resultados do *mining*, para melhorar e compreender o comportamento do cliente. Mas contudo para retirar o máximo potencial das previsões estas devem ser transferidas para o domínio das operações. Os produto de *Blue Martini* são únicos nas facilidades de integração e transferencia de regras e associações do *data mining* para o ambiente de operações. O motor de *mining* pode gerar e transferir cross-sells automaticamente. Sem esta arquitectura de integração, as personalizações em massas seriam impraticáveis.

5.2.1 **Características do sistema**

Gestão de Campanhas

- ✗ Campanhas direccionadas para os clientes, e futuros clientes, e associados
- ✗ Campanhas com multiplos objectivos, promoções, e newsletters,
- ✗ Anuncios, pesquisas, eventos, convites e grupos virtuais.
- ✗ E-mail, e Webmail
- ✗ Inclusão automática de grupos de controlo.
- ✗ Relatórios pre-construidos (ex. tempode resposta e ROI)
- ✗ Lista de gestão.

Armazéns de dados

- ✗ Perfil de Clientes
- ✗ Schema optimo para analise
- ✗ Multi-objectivo ETL (*Extract, Transform, Load*)
- ✗ Clickstream and business event logging
- ✗ Integração online com *Acxiom Data Network*
- ✗ Bases de dados Oracle ou Microsoft SQL

Relatórios

- ✗ Mais de 50 relatórios pre-definidos como resposta a uma campanha, ROI, vendas sales, *clickstream*, produtos mais vendidos, produtos menos vendidos, falhas frequentes em pesquisas.
- ✗ Relatórios Ad hoc
- ✗ Representações por tabelas e gráficas
- ✗ Updates automáticos

Centro de analise baseada na Web

- ✗ Acesso por browser
- ✗ Repositório unificado para relatórios, visualizações, *data mining*, investigações, e ferramentas OLAP

Data Mining

- ✗ Multiploa algoritmos: C5.0, Associações, *GritBot*
- ✗ Motor de Indução de regras
- ✗ Sugestões *cross-sell*
- ✗ Análise estatística

Visualização dos dados

- ✗ Tipos de visualização: 5D *scatterplots*, explorador em árvore, series de tempo, filtros para gráficos, histogramas, gráficos de barras, e tabelas
- ✗ Formato interactivo

Internacionalização

- ✗ Suporta qualquer linguagem via *Unicode Standard*
- ✗ suporta multiplas moedas
- ✗ Interface em inglês, francês, alemão, japonês, e espanhol

Personalização

- ✗ personalizaçãode: conteúdos, imagens, produtos, preço, promoções
- ✗ Multiplos pontos de contacto: *website*, *call center* e loja
- ✗ tecnologia *wireless*
- ✗ Identificação automatica de regras utilizando as tecnicas de *data mining*: regras if-then e associações
- ✗ Implementação automatica de *cross-sells*
- ✗ Entrada manual (optional) em Inglês
- ✗ Gerenciador de regras e de conjuntos de regras
- ✗ Filtra o to bloco de regras inapropriadas ou ofertas para quais não existe *stock*

Transformação dos dados

- ✗ Transformações de dados: agregação, adicionar add colunas, apaga colunas, filtra, ordena, e une os dados

Limpeza dos dados

- ✗ Detecção de anomalias

Workflow

- ✗ Criação de tarefas
- ✗ Escalonamento de tarefas
- ✗ Redireccionamento de tarefas

Infraestrutura

- ✗ Aplicação de servidor
- ✗ Compatível com J2EE
- ✗ Arquitectura *three-tier*
- ✗ *Java Server Pages*

Integração

- ✗ Integração visual: EJB, XML, Microsoft COM/DCOM, CORBA, e MQ Series da IBM
- ✗ Adaptadores pre-construídos: STC/SeeBeyond Eway Adapter, SAP, Siebel

Segurança

- ✗ Utilizadores com diferentes privilégios
- ✗ LDAP (*Lightweight Directory Access Protocol*)

Fácil de utilizar

- ✗ Ferramentas divididas por módulos
- ✗ Interface gráfico

Especificações de sistema

Servidores que suporta

- ☞ Sun Solaris 8
- ☞ Microsoft Windows NT Server 4.0
- ☞ HP-UX 11.0

Base de dados que suporta

- ☞ Oracle ORACLE 8.1.7
- ☞ Microsoft SQL Server 7.0

Servidores *Web* que suporta

- ☞ Microsoft Windows NT 4.0 and 5.0 (Windows 2000)
- ☞ iPlanet Web Server 4.1 (inclui suporte para Netscape Enterprise Server)

Dispositivos *mobiles* e *sem fios*

- ☞ Web-enabled Cellular Phones (WAP & i-Mode)
- ☞ Palm Vllx
- ☞ Symbol Mobile Wireless Handheld Devices

5.3 Clementine



<http://www.spss.com>

Oferece sequências de associações e *clustering* usado para análise de dados da Web.

No dia a dia de qualquer organização são produzidas grandes quantidades de informação. Esta informação retém – e por vezes esconde – a experiência passada. Analisar esta informação pode resultar numa maior capacidade de decisão e na melhoria da relação com os nossos clientes.

Com o Clementine pode aprender com o passado para melhorar o seu futuro. O Clementine é uma ferramenta de *data mining*. Descobrir padrões e tendências que nos permitam trabalhar mais eficientemente é um dos objectivos do *data mining*.

O conjunto exaustivo de técnicas disponíveis no Clementine permite encontrar respostas para os desafios analíticos mais extremos. Começando por combinar dois dos elementos mais preciosos – os nossos dados e o conhecimento que temos do negócio – podemos criar poderosos modelos de negócio.

Desde redes neuronais a árvores de decisão e técnicas de regressão logística até procedimentos de apresentação e preparação dos dados, o Clementine pode utilizá-los para criar os melhores modelos explicativos em áreas tão diversas como:

- Prever o comportamento dos clientes, analisar o risco e detectar excepções, de modo a prestar-lhes um melhor serviço.
- Analisar os dados de acesso ao seu portal *Internet (eMine)* para descobrir o percurso dos visitantes e, com isso, personalizar e rentabilizar os acessos.
- Classificar os clientes em categorias específicas, tais como rentabilidade alta ou baixa e potencial de retenção alto ou baixo, para melhor definição de estratégias.
- Identificar casos de fraude ou de não conformidade, para evitar a diminuição de receitas
- Prever vendas ou utilização de serviços de modo a aplicar os recursos disponíveis onde vão ser mais necessários.
- Identificar grupos similares para objectivos de personalização
- Executar análises tipo '*market-basket*' para descobrir que tipos de produtos ou serviços são comprados em conjunto.

A arquitectura do Clementine facilita o processo de descoberta.

O processo de análise é claro. A aproximação interactiva ao *data mining* através de 'streams' é a chave do processo. Utilizando ícones que representam passos no processo de *data mining*, analisamos os nossos dados construindo uma 'stream' – um mapa visual do processo através do qual fluem os dados. Comece por colocar um ícone de fonte dos dados, arrastando-o da conjunto de objectos presente no écran na área de acesso a dados. A seguir explore-os visualmente através de gráficos para conhecê-los. Aplique um ou mais dos algoritmos disponíveis para construir o seu modelo. O seu 'stream' está construído, e o que fez foi pura e simplesmente colocar vários ícones sequencialmente, ligando todo o processo, a partir de um conjunto de objectos que lhe é apresentada no ecrã.

Os 'streams' são interactivos. Cada passo do 'stream' é seleccionado com base em dois princípios – a informação que obtivemos na etapa anterior e o nosso conhecimento do negócio. Deste modo podemos adaptar continuamente o nosso 'stream'.

Os modelos são facilmente criados e testados. Pode construir e testar múltiplos modelos para ver imediatamente qual o modelo que melhores resultados produz. Ou pode combiná-los, utilizando os resultados de um modelo como input de um outro modelo. Estes 'meta-modelos' tomam em conta as decisões do modelo inicial e podem melhorá-los substancialmente.

Visualizar os resultados permite-nos compreender as alterações ocorridas. As poderosas técnicas de visualização do Clementine permitem-nos compreender as relações importantes nos dados e determinar o melhor caminho de análise a prosseguir. Através de gráficos interactivos podemos ver os padrões dos dados em qualquer passo do processo. E através da função 'query by mouse' podemos explorar esses padrões seleccionando subconjuntos dos dados directamente a partir do gráfico.

A PSE está apta a analisar consigo o seu problema e construir protótipos que demonstrem os ganhos efectivos decorrentes da adopção da solução Clementine para um processo de *data mining* da sua informação.

5.3.1 Clementine 6.5

Clementine 6.5 traz novos recursos que agilizam análise e exploração de dados em aplicações voltadas para relacionamento com cliente

A SPSS, líder em tecnologia de *data mining* e CRM analítico, lançou um *software* de *Data mining* voltado para gestão de relacionamento com clientes. O Clementine 6.5 traz como diferencial uma série de recursos baseados em templates (CATs – *Clementine Application Templates*) que ajuda os utilizadores a montar passo-a-passo o mapeamento dos clientes, bem como elaborar projecções de negócios. O sistema permite criar modelos que indicam migração de clientes, que produtos e serviços são mais consumidos e utilizados em conjunto, as páginas e sequências mais consultadas dentro do *Website* e outras informações para incrementar as soluções de CRM.

O Clementine 6.5 mantém as funções básicas do *software* original. O programa ajuda a desenvolver modelos de previsão inclusive gráficos, indexados a cada questão específica do negócio, identificar e classificar grupos de consumidores de acordo com perfil e comportamento, faz agrupamento de casos, entre outros. Com base nessas funcionalidades, os utilizadores podem traçar o perfil do cliente, identificar oportunidade de pré e pós venda, atingir novo público alvo, detectar fraudes, reduzir riscos de negócios e aprimorar serviços e comércio electrónico.

No entanto, a nova versão do *software* reúne uma série de funções e interface mais interactiva que facilitam e agilizam todo trabalho analítico. O CRM CAT, por exemplo, ajuda a aumentar o *lifetime value* de um cliente, revelando informações que contribuem para prever taxas de resposta de acções promocionais.

O *software* agora conta com a função *Operation Reordering*, para reordenar automaticamente as operações, de modo a enviar quantidade de tarefas do Clementine para a própria base de dados, o que resulta em melhoria do desempenho e diminuição do tráfego na rede. Outra novidade é o *Expression Builder*, sistema que permite seleccionar funções de operação, selecção de campos e manipulação de dados a partir do mouse, o que minimiza erros no processo e o trabalho de digitar expressões complexas. A ferramenta também faz subdivisão de grupos de dados de acordo com características peculiares e apresenta um novo algoritmo sequencial para identificar padrões de informações em ordem cronológica e fazer previsões sequenciais.

O Clementine 6.5 chega com outras funções aperfeiçoadas. É o caso da *Web mining* CAT, usada para colher dados e fazer projecções sobre actividades do *Web site*. O sistema vem com quatorze novos modelos de mapeamento (*streams*), com espaços para registrar logs comuns, estendidos e *Microsoft IIS*. O dispositivo pode ser usado para agregar partes do site no modelo de previsão, registrar o tempo de navegação do cliente, as páginas mais consultadas e serviços Web mais procurados.

Cada vez mais as empresas estão constatando que as ferramentas analíticas são fundamentais para o sucesso do CRM. E o lançamento da edição 6.5 vem reforçar a posição do Clementine como a mais prática e completa tecnologia de *data mining* para gestão de relacionamento com clientes. O Clementine já é utilizada por mais de 500 companhias no mundo, dos segmentos de manufactura, telecomunicações, finanças, varejo, saúde, governo, universidades e *e-commerce*.

O produto é compatível com o sistema operacional Windows 95, 98, 2000, XP, ou NT4.0 (cliente); Windows 2000, NT4.0, Solaris 2.6, 7 ou 8, HP/UX 10.20 ou 11, AIX 4.2.2 ou 4.3 (Server). Disponível na versão cliente-servidor.

5.3.2 SmartViewer Web Server

O *SmartViewer Web Server* é uma solução para difundir os resultados analíticas pelos administradores, departamentos ou clientes a que se destinam.

Com esta solução é-lhe possível distribuir os resultados analíticos obtidos com o SPSS de um modo fácil e num meio de acesso generalizado – através da *Internet* ou da sua intranet e por utilização do *Web browser* que utiliza para outros fins.

O *SmartViewer Web Server* torna fácil e rápida a publicação dos resultados SPSS e a sua distribuição a indivíduos no interior ou fora da sua instituição. Uma vez publicados, é-lhe igualmente permitido actualizá-los reflectindo de imediato qualquer alteração aos dados originais.

O processo de tomada de decisões torna-se mais fácil. Em primeiro lugar o analista cria um relatório com base no SPSS (gráficos ou quadros de resultados). De seguida – tal como arquiva um ficheiro – o analista publica o seu relatório, inserindo-o no *SmartViewer Web Server*, tornando-o acessível. O potencial receptor, através do *Web browser* do seu computador, pode a partir de agora aceder ao catálogo que estrutura a informação disponibilizada e visualizar os relatórios que pretende.

Este acesso não é público.

Com o *SmartViewer Web Server* o acesso à informação pode ser protegido e seguro. O Administrador da solução tem controlo sobre o acesso à informação. Pode especificar que documentos estão acessíveis para o receptor X através da atribuição de privilégios de acesso ou de restrições nos destinatários

Exemplos de aplicação

➤ **Comércio.** Uma cadeia de lojas está interessada em abrir mais estabelecimentos. A sua localização potencial foi objecto de análise. O *SmartViewer Web Server* é utilizado para distribuir a informação obtida à cadeia de gestão da empresa localizada em diferentes lugares. Cada um dos administradores com base na informação recebida pode fazer os seus comentários num documento *Word* e voltar a distribuí-lo aos seus parceiros através do *SmartViewer Web Server*.

➤ **Administração Pública.** Os novos estilos de governo obrigam qualquer departamento governamental a disponibilizar informação analítica – por exemplo, estudos demográficos, execução de programas, etc. – para acesso ao público em geral. O *SmartViewer Web Server* é um instrumento ideal para cumprir esta função, mais que não seja porque através da difusão de tabelas interactivas permite a cada cidadão uma visão personalizada da informação.

➤ **Serviços Financeiros.** Uma das principais empresas gestoras de cartões de crédito investe muito dinheiro em campanhas de *direct mailing* para atrair novos clientes. Com o SPSS determina os seus melhores alvos. Com o *SmartViewer Web Server* reporta os resultados das campanhas em curso. Os gestores de cada produto têm acesso imediato ao grau de sucesso de cada campanha que gere.

➤ **Telecomunicações.** Um operador tem em vigor três planos de

comercialização distintos e utiliza o SmartViewer Web Server para guardar, distribuir e reportar internamente informação vária sobre cada um dos planos. Os gestores acedem a essa informação para melhor decidir sobre as acções de *marketing* a tomar.

- **Saúde.** Num Hospital cada um dos serviços clínicos tem em vigor um sistema de recolha de informação sobre os níveis de satisfação dos seus doentes. Após análise dos resultados, são produzidos relatórios individuais e agregados que chegam á Administração através do SmartViewer Web Server.
- **Estudos de Mercado.** Uma empresa internacional produz estudos de imagem sobre diversas marcas. Mediante contratos com os seus clientes, esta empresa disponibiliza a cada cliente a informação vendida através do SmartViewer Web Server.

5.4 Megaputer WebAnalyst



<http://megaputer.com>

WebAnalyst (WA)

WA é um servidor analítico que permite a integração dos armazenes de dados com as técnicas de *data mining*, com ênfase no e-business e aplicações de *Web mining*. WA é uma aplicação escalável de servidor com uma arquitectura aberta que torna autónomas as tarefas de e-business de recolha de dados, de transformação, análise e personalização das interações com os clientes. O WA permite um ambiente visual de programação para o analista.

WebAnalyst (WA) é uma aplicação de servidor que permite:

- Processa dados de fontes diferentes, como os canais da *Web* (HTTP), de bases de dados externas, e de ficheiro de acessos ao servidor bases de dados, ficheiros de registo de acessos.
- Guarda todos os dados numa base de dados unificada do WA
- Contém uma ferramenta de processamento de dados
- Permite ao utilizador um ambiente visual de programação e gera procedimento analíticos reutilizáveis.

Objectivos

WebAnalyst ajuda:

- Guarda todas as interações do cliente
- Transforma e guarda os dados de num formato conveniente para uma análise futura
- Utiliza os dados para aprender sobre todos os interesses, preferencias do cliente
- Analisa os recursos e a arquitectura do *website*
- Gera relatórios
- Reconhece os clientes mais frequentes e acede ao seu perfil
- Recolhe informação para personalizar as comunicações com os clientes

Os recursos de um WA:

- Web Server
- Conteúdo da base de dados
- Base de dados de produtos e transacções
- Ficheiros de registo de acessos

As tarefas executadas pelo WebAnalyst

WA pode executar enumeras tarefas, tarefas estas que estão divididas e sete grupos:

1. recolhe a informação dos diferentes canais da *Web* (HTTP, SMTP, NNTP, e FTP).
2. Processa os ficheiros de registo de acessos do servidor.
3. Executar diferentes tarefas nos armazéns de dados: extrai informação das base de dados externas, executa limpeza aos dados guarda no servidor do WA.
4. Executas tarefas de processamento analítico com a ajuda do explorador do PolyAnalyst e com os módulos do *data mining*.
5. Executa vários objectos de WASL: Existentes já no WA ou criados pelos utilizadores .
6. retorna a informação ao canal de transmissão gerando conteúdos e relatórios analíticos
7. Permite ao utilizador um ambiente gráfico de programação

Arquitectura do servidor

O componente principal do WA server é o seu Sistema. WA server pode ser dividido em três componentes: Transaction Manager, Channel Processors Manager, e Virtual Machine.

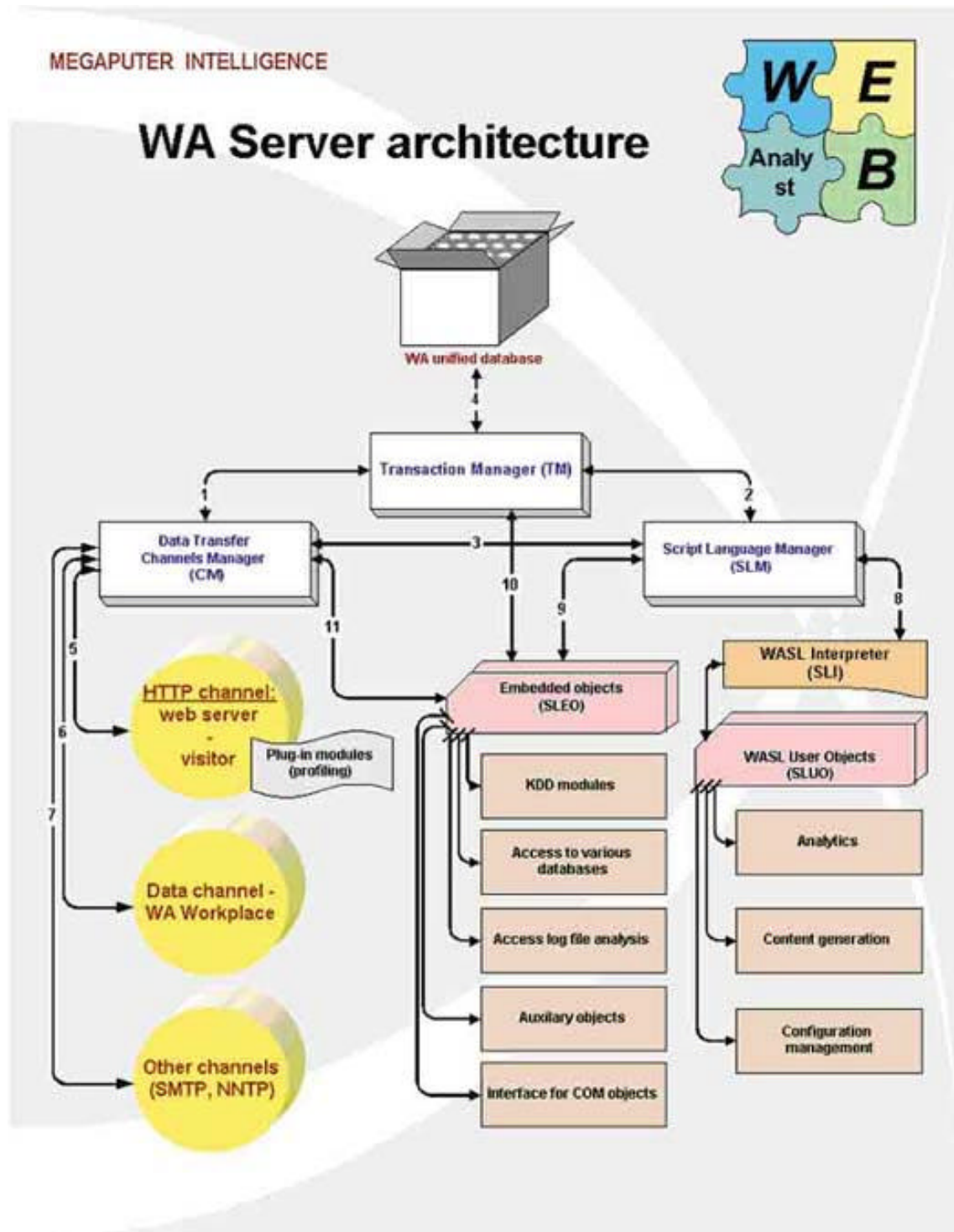


FIGURA 21: ARQUITECTURA DO WEBANALYST

5.5 MicroStrategy Web Traffic Analysis Module



<http://www.microstrategy.com>

5.5.1 Domínio da Análise

Web Traffic Analysis é parte integrante de uma estratégia de negócio on-line. Através do *Web Traffic Analysis*, podemos medir e analisar as interações dos clientes na Web. Estas interações podem ocorrer em múltiplos processos de negócio tais como: vendas, prestação de serviços, e marketing, etc.

5.5.2 A solução do Modulo de Análise de trafego da MICROSTRATEGY

O modulo de análise de trafego da MicroStrategy, esta incluído na plataforma MicroStrategy 7 que permite a análise de terabytes de trafego da Web e também os dados dos utilizadores. O modulo contém 64 *packaged reports* com a possibilidade de estes serem alterados. O modulo de análise de trafego alterado para conter de forma a conter outros reports ou outra facilidade para a análise de dados

5.5.3 Áreas chave da análise

Foi desenhado para ser modular. O modulo de análise de trafego da Web facilita a obtenção de insight em vários processos dentro da Web. O modulo permite a sumarização de detalhes nos relatórios que contem.

Os relatórios são divididos em três grandes áreas:

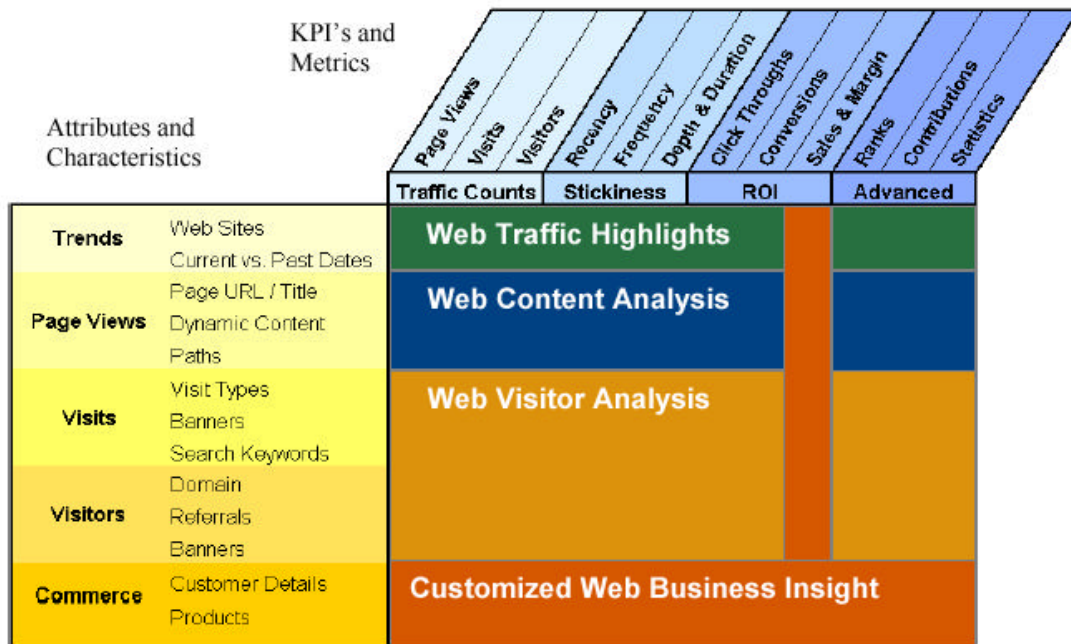
- *Web Traffic Highlights* – 10 reports
- *Web Content Analysis* – 22 reports
- *Web Visitor Analysis* – 32 reports

5.5.4 Organização da análise

As três áreas de análise tem um número de relatórios associados. Estes relatórios contem atributos e indicadores de desempenho (KPI's) que se encontram pré-construídos em MicroStrategy 7.

Os dados podem ser divididos e serem vistos por perspectivas diferentes como: produtos, conteúdo, e características dos utilizadores permitindo identificar os pontos fracos do negocio.

O gráfico seguinte ilustra os atributos e características em conjunto com as métricas do negocio e as KPI's que servem como blocos à análise. Estes blocos podem ser combinados para produzir mais dos 64 relatórios que vem no modulo.



5.5.5 Web Traffic Highlights

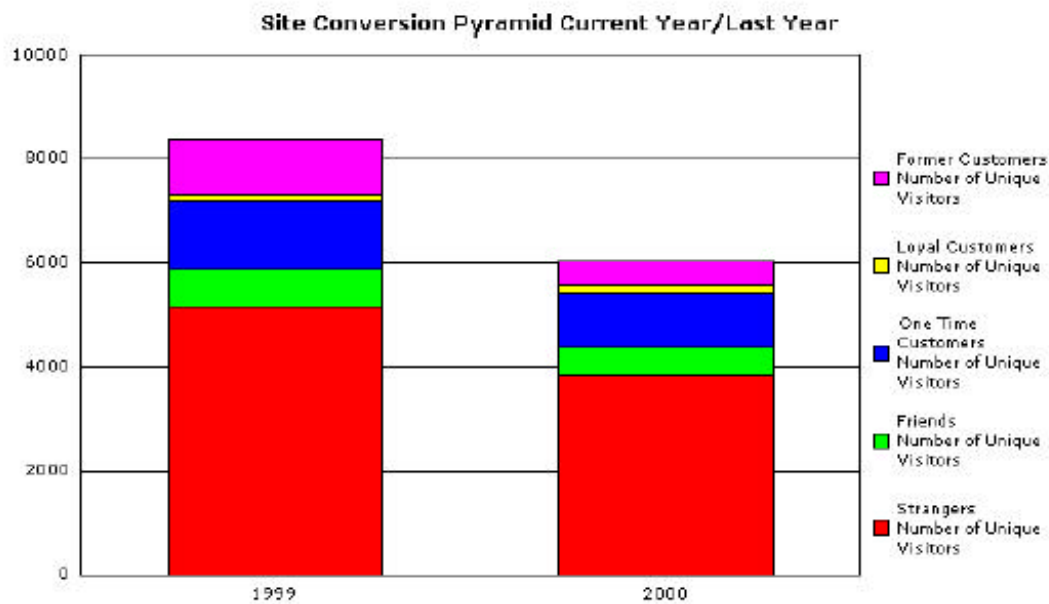
Web Traffic Highlights permite analisar as tendências e alertas nos *web sites*. Com comparação baseadas no tempo, indica se as alterações no tráfego da *Web* são ou não sazonal. Ao alertas são dadas através email ou outro meio. Estes relatórios levam a investigação noutras áreas.

As perguntas que este modulo permite responder:

- Quais as tendências nos *website* em diferentes períodos de tempo como este ano vs o ano passado.
- Existem alterações anormais nas tendências sazonais que precisão de ser investigadas.
- Quais os cenários de negocio que permitem um resultado positivo.

5.5.6 Web Visitor Analysis

Web Visitor Analysis analisa as tendências de comportamentos quer de utilizadores anónimos quer utilizadores registrados também inclui características dos utilizadores e a frequência das visitas. Esta técnica de análise pode ser utilizada para medir os efeitos de uma estratégia para melhorar a atracção, conversão e retenção de clientes.



Perguntas que permite responder:

- Quais os *banners* de publicidade e outros tipos de *links* são visitados ou não pelos clientes registrados?
- Á percentagem de sucesso em converter utilizadores em utilizadores frequentes e em clientes?
- Quais os clientes que estão a ser perdidos?

5.5.7 Web Content Analysis

Web Content Analysis permite monitorizar o conteúdo estático e dinâmico de um *website*, conteúdo esse que atrai, converte, retém os utilizadores da *Web*. Isto permite otimizar o conteúdo, o layout e a navegação do *website*. Os utilizadores do modulo podem seguir a actividade de um cliente específico como as palavras chave utilizadas e actividade de compras

Perguntas que permite responder:

- Como é que os utilizadores que visitam o site pela primeira vez o utilizam vs. Os utilizadores frequentes
- Como se optimiza a sequência de páginas no *site* para ter as melhores taxas de conversão de clientes?
- Quais as páginas que fazem o cliente desistir do *site*?

5.6 XML Miner



<http://www.metadatamining.com/>

5.6.1 Data Mining Package

Analisa, filtrar, classifica, e prevê dados de XML. XML Miner analisa extrai conhecimento dos dados de XML. Utiliza para isso a tecnologia de Fuzzy Rule induction, que gera regras que explicam e prevêem valores seleccionados como *input* no conjunto de dados, baseados em outros valores. Resultando num conjunto de regras expresso em Meta-regras, um dialecto do XML que pode ser convertido em inglês utilizando o XSL, e também é compreendido pelo processador de regras do XML.

5.6.2 XML Miner

- A versão de *single-user* permite uso ilimitado num único computador sem tempo limite, mas não funciona em servidores *Windows*.
- A versão de servidor permite uso ilimitado num servidor sem tempo limite.

Os componentes do XML Miner constituem um *kit* completo a extracção de dados e de fontes de dados em XML, gerando as regras nos termos de regras *if..then..* e regras fuzzy e em XML baseado no conhecimento representação da linguagem, Metaregras, e regras em inglês em HTML e usando regras de descoberta no mundo real no *Web site*, um *browser*, ou em qualquer aplicação *Windows*.

O XML Miner também inclui um objecto, *strucfind*, que é uma árvore de controlo que extrai o *schema* tipo de informação dos dados do XML e mostra a estrutura em forma de árvore. Usando este objecto o utilizador pode facilmente seleccionar o *input* e o *output* a ser utilizado para a extracção de conhecimento.



Existem varias metodologia para a extracção do conhecimento dos conjuntos de dados. XML Miner usa um algoritmo de indução de regras que gera resultados compreensíveis e utilizando o XSL transforma as regras em inglês que permite explicar exactamente o que foi descoberto e os *inputs*, *outputs*, *fuzzy sets*, e as categorias derivadas dos dados.

XML Miner permite o utilizador especificar a percentagem dos dados que são utilizado para o conjunto de teste, e depois da indução das regras o XML Miner calcula o desempenho dos dados teste e de treino.

XML Miner trata tipos de dados primitivos, e classifica *inputs*, automaticamente converte em tipos categóricos e numéricos. O XML Miner classifica dados ou prevê valores numéricos.

XML Miner recebe como *input* a URL da fonte de dados ou uma *string* de XML, e instruções como a previsão deve ser baseada e em que itens. Existem duas formas de instruções: um XML baseado na definição do ficheiro, (como um gerado pela utilidade *strucfind*) ou os *inputs* e *output* podem ser individualmente especificados em Visual Basic como as colecções onde o XPath como outros paramentos podem ser especificados.

O resultado da análise são Metaregras que descrevem o conjunto de dados, e dois valores descrevendo o desempenho dos dados de treino e de teste. Para os dados categóricos o *output* do desempenho é expresso se a fracção foi classificada correctamente.

Requisitos de sistema:

- Windows 9x, 2000, ME, XP and NT
- Espaço em disco: 6MB
- Memória: 32MB
- Software: Microsoft Internet Explorer 5.0

5.7 WebTrends



5.7.1 WebTrends Log Analyzer

Log Analyzer é um *software* de análise de tráfego da *web* destinados a administradores da *web* e de *marketing*. É ideal para os pequenos negócios permite obter informação útil sobre o comportamento dos utilizadores e ajuda a melhorar o desempenho do *site*.

Configurado para um único servidor de *web* de *site*, *Log Analyzer* produz essencialmente relatórios sobre os padrões dos visitantes, referentes ao seu comportamento, dados demográficos.

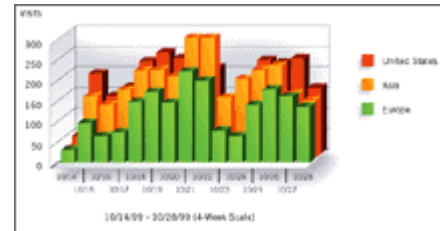


FIGURA 22: ESTATISTICAS

Log Analyzer representa os dados de forma organizada através de gráficos colorido. Tendo maior ênfase em tornar a informação fácil de ler e de interpretar

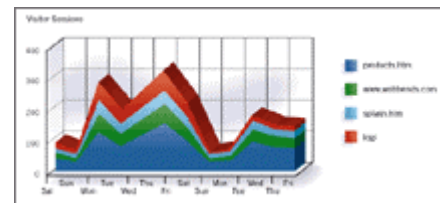


FIGURA 23: MODO GRÁFICO

Log Analyzer v7.0 Características

- Instalação com *Wizards* e ajuda através do *Professor WebTrends tutorial* e funções de ajuda.
- Permite um rápido acesso ao perfis dos visitante e contém 15 relatório pre-formatados para uso imediato.
- Expansão para 50 relatórios de perfis e permite a customização de formatos de relatórios.
- Permite a programação de funções automáticas de geração de relatorios bem como o processamento de ficheiros de batch.
- Fácil reanálise de relatorios de perfis com a base de dados de FastTrends® e melhora o desempenho da arquitectura de programas C++.

5.7.2 **WebTrends Intelligence Suite**

NetIQ's *WebTrends Intelligence Suite*, permite uma vista completa da actividade do cliente para poder medir a melhoria do desempenho.

A *web* tornou-se uma parte intrínseca do dia a dia dos negócios, e cada vez se torna mais crescente a necessidade de medir o retorno do investimento de *e-business*.

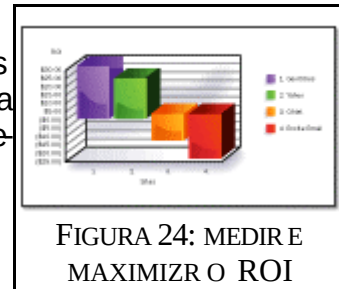


FIGURA 24: MEDIR E MAXIMIZAR O ROI

WebTrends Intelligence Suite inclui:

- **WebTrends Warehouse.** Transforma, processa, e guarda todos os detalhes da actividade do cliente da *Web* para uma posterior análise.
- **WebTrends Reporting Center (Enterprise Edition).** Contém mais de 350 gráficos e tabelas predefinidas.
- **WebTrends Report Designer.** Permite desenhar e customizar relatórios, e permite a análise comparativa de relatórios
- **WebTrends OLAP Manager.** Contem mais de 40 modelos multi-dimensionais pre-definidos, e permite criar novas análise para analisar novas tendências.
- **WebTrends para Administração de Sistemas.** Traduz e integrar formatos proprietários para a análise.

Características WebTrends Intelligence Suite

- **Máximo retorno de investimento no *web site*.** Medindo o sucesso da forma de disseminação de informação a forma como o cliente de *e-commerce* age, isto permite avaliar o que funciona e o que não funciona no *site*.
- **Melhora a eficácia do *marketing*.** Permite identificar como é que os tráfego de visitante chega ao site, e que clientes respondem as campanhas e quais as campanhas que são mais eficazes em trazerem clientes.
- **Melhora a usabilidade do *Web site*.** Analisa quais os clientes que passam da página de entrada, quais os elementos que são lentos ou que quebram a sessão do cliente e quais os erros técnicos que levam a um decréscimo das transacções.
- **Analisa o desempenho do *e-commerce*.** Segue todas as ordens de venda e o desenvolvimento dos produtos e analisa as tendências de

visitantes que passam a clientes, incluindo a actividade do cesto de compras.

- **Contém relatórios múltiplos e níveis de análise.** Das estatísticas básicas do *web site* as multi-dimensionais análises *slice-and-dice*, relatórios pre-definidos e customizados que dão informação sobre os utilizadores, tendências de tráfego, ajustar as campanhas de marketing e previsão e cálculo do ROI.
- **Recolha de dados da web data, e administração dos dados guardados.** *WebTrends Warehouse* permite importar e transformar a transacção individual de um visitante em informação útil. Oferece muitas maneiras de identificar os visitantes e de classificar os dados da web de forma de identificar acções ou eventos.
- **Integração com infraestrutura existente.** Para uma análise completa e precisa, *WebTrends Intelligence Suite* permite integrar os dados da web com o sistema existente de administração, sistema *e-commerces*, CRM, ERP. Integração com múltiplas bases de dados, plataformas e *browsers*.



<http://www.123loganalyzer.com/>

5.8 123LogAnalyzer

123LogAnalyzer permite ver o perfil dos utilizadores:

- Como navegam no *website*
- Quais as páginas que são vistas, quais as que são ignoradas
- Quanto tempo ficam no site
- De onde é que vieram
- Quais as novas oportunidades para atrair mais visitantes/clientes
- Se existe uma mudança no comportamento de um visitante
- Se alguém está a fazer algo de errado no *Web site*



123LogAnalyzer pode analisar ficheiros ZIP and GZ com os ficheiros de registo de acessos sem ter que os descomprimir primeiro.

Outro *software* de análise de ficheiro de registo de acesso utilizam o *reverse DNS lookup* tecnologia que gera informação geográfica, que é altamente imprecisa. 123Loganalyzer usa um IP interno para mapear, tecnologia que identifica a região geográfica do visitante pelo seu endereço de IP.

123LogAnalyzer permite ver como é que as pessoas utilizam os motores de pesquisa para encontrar um determinado site. O relatório de desempenho do motor de pesquisa permite saber quais as palavras utilizadas nas pesquisas, quais os motores de pesquisa que tem maior (ou menor) tráfego. Estes relatórios permitem otimizar as páginas da *Web* e ajustar as campanhas de marketing para ter mais tráfego vindo dos motores de pesquisa. 123LogAnalyzer dá um relatório com os detalhes do visitante/IP com o tempo de acesso, *hits* e *bandwidth usage*. 123LogAnalyzer dá um relatório detalhado dos *links* partidos e as páginas com erro no site.

A figura seguinte mostra o ecrã de entrada do 123LogAnalyzer esta ferramenta permite obter resultados rápidos a partir de um ficheiro de log

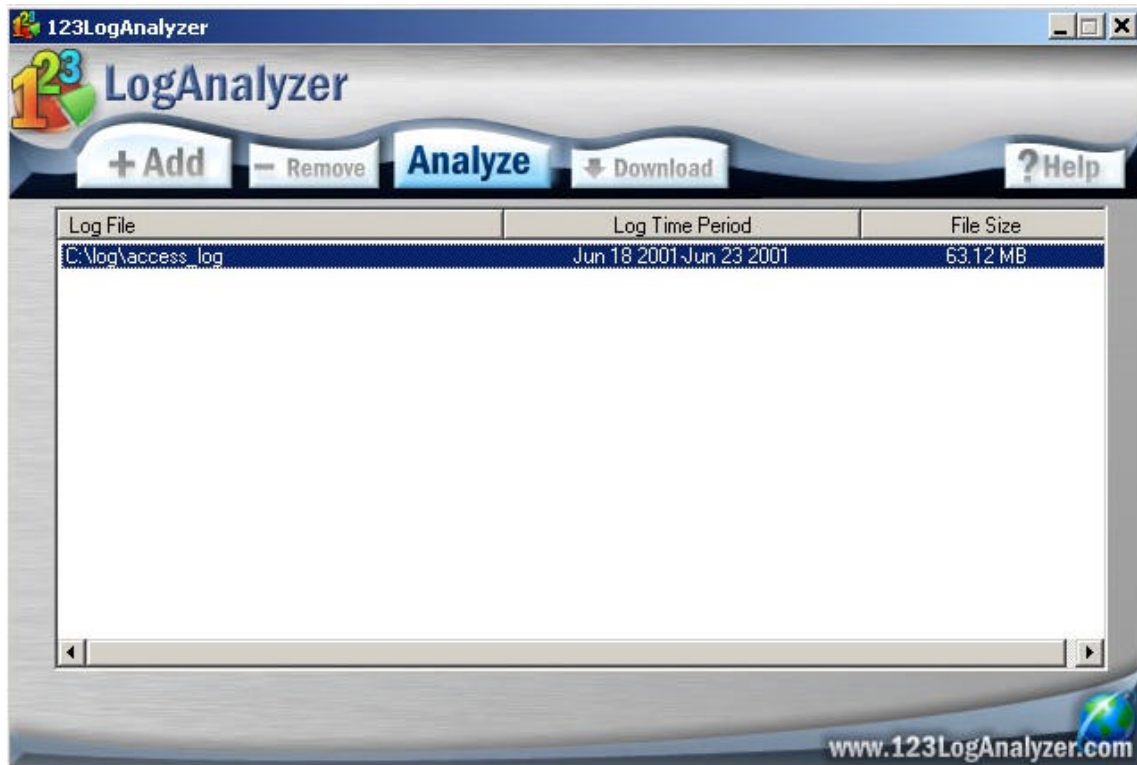


FIGURA 25: 123LOGANALYZER

O botão Add: Abre uma janela que permite escolher o ficheiro(s) de log. Este pode estar no seguinte formato .log, .zip, .gz, ou .txt.

O botão Remove: Remove o ficheiro de log seleccionado

O botão Analyze: Abre uma janela que permite introduzir o domínio, amplitude das das, a utilização de filtros, e as preferencias de output. Quando o 123LogAnalyzer acaba lde analisar o ficheiro automaticamente gera filheiros de relatóriopodendo estes serem vistos num *web browser*.

O botão Download: Abre uma janela que permite fazer *download* do ficheiro de *log* de um servidor de FTP ou de um servidor Web. .

O botão Help: Abre uma janela de ajuda.

Relatórios

123LogAnalyzer gera automaticamente relatórios cada vez que analisa um ficheiro de *log*. Os relatórios são bastantes extensivos, com a informação num formato fácil de ler. O relatórios são em formato HTM.

Os relatórios estão divididos por categories:

Estatísticas gerais**Actividade Estatísticas**

- ↗ por incremento de tempo
- ↗ por dia da semana
- ↗ por hora do dia

- ↗ por referencia do dominio
- ↗ por referencia às URLs

Estatísticas do motor de pesquisa

Pesquisas

- ↗ por palavra chave
- ↗ por motor de pesquisa

Recurso Acedido

- ↗ por *page views*
- ↗ por sequências de *browsing*
- ↗ por ficheiros *downloaded*
- ↗ por directorios acedidos
- ↗ por imagens acedidas

Região geográfica**Browsers e plataformas****Informação técnica**

- ↗ Páginas não encontradas (404)
- ↗ Erros em geral

Referencias estatísticas**Requisitos do sistema**

Microsoft Windows 95/98/2000/NT/ME/XP
Pentium 90mhz ou acima
64 MB RAM
10 MB de espaço de disco

5.9 Webanalyzer

The Webalizer

What is your web server doing today?

<http://www.mrunix.net/webalizer/>

O Webalizer é um programa de análise de logs. Produz relatórios em formato HTML, que podem ser vistos em qualquer *browser*.

O Webalizer produz estatísticas anuais, mensais, diárias e horárias. Nos relatórios mensais diversas estatísticas são produzidas sobre as utilizações diárias, horárias, pelos sites visitados, URL, *browsers*, totais de página e visitas, os totais das entradas e saídas das páginas, *string* de pesquisa, etc. As estatísticas produzidas pelo Webanalyzer são:



➤ Acessos (*Hits*)

Qualquer pedido ao servidor e é registado no ficheiro de registo de acessos é considerado um acesso. Os pedidos podem ser de páginas html, gráficos, imagens, ficheiros de audio, scripts de cgi, etc... O número total de pedidos representa o número de pedidos feitos ao servidor em determinado período.

➤ Ficheiros

Alguns dos pedidos feitos ao servidor requerem que seja mandado ao cliente, por exemplo uma página html ou uma imagem. Quando isto acontece, é considerado um ficheiro e o total de ficheiros é incrementado. A relação entre os acessos e os ficheiros pode ser descrita em termos de pedidos que chegam (*incoming requests*) e respostas que saem (*outgoing responses*).

➤ Páginas (*Pageviews*)

Geralmente um documento HTML ou algo que gera um documento em HTML é considerado uma página. Isto não inclui os componentes das páginas como: imagens, ficheiros de audio, etc... Este número representa o número de páginas pedidas.

Sites

Cada pedido feito ao servidor vem de um único *site*, que pode ser referenciado pelo nome ou um endereço IP. O número de *sites* mostra quanto endereços IP podem fazer pedidos ao servidor durante o período em que se realiza a análise. No entanto este número não corresponde ao número de pessoas individuais que visitaram o *site*.

Visitas

Quando um pedido, um endereço IP (*site*), ao servidor, o tempo que demora desde do ultimo pedido é calculado e se for maior do que o tempo configurado aparece o valor de *visit timeout*, (ou nunca foi pedido antes), é considerado como uma visita nova e o total é incrementado (tanto os número de *sites* e de endereços de IP).

KBytes

O valor dos KBytes (kilobytes) mostra a quantidade de dados, em KB, que foi enviado para for a do servidor durante o período da análise.

As páginas mais utilizadas para entrada e saída do servidor

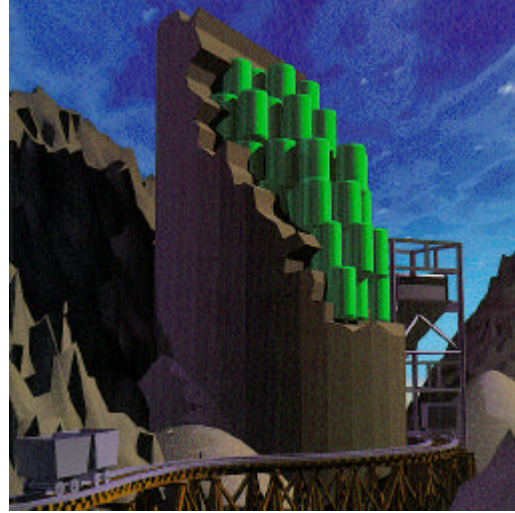
As páginas mais utilizadas para entrada e saída do servidor dão uma estimative de quias as URLs utilizadas para entra no site e quais as páginas do *site* que causaram a saída do *site*

Características

-  O Webanalyzer foi constroido em C o que o torna rápido e protavelto be extremely fast and highly portable.
-  Suporta os *standards* dos formato dos ficheiro de registo de acessos bem como outras variações deste tipo de ficheiro. Permite gerar estatísticas referindo o tipo de *sites* e *browsers* e também supotas formatos de ficheros de log do wu-ftp d xferlog FTP e squid.
-  Gera relatórios que podem ser configurados apartir da linha de comandos ou utilizando ficheiros de configuração
-  Suporta multiplas linguagens.
-  Tamanho de ficheiro de registo de acesso pode ser ilimitado
-  É distribuido pelo GNU General Public License, e o código completo está disponivel, bem como as distribuições binárias.

Conclusão

Quando um sistema de extracção de conhecimento é bem planeado, e está baseado em dados fiáveis, e é bem absorvido pela instituição, tem-se uma das melhores previsões das acções de um cliente, que é o seu histórico de comportamento e compras realizadas. O mercado de transacções ao nível da *web* torna-se o cenário ideal para extracção de conhecimento. Diariamente são adicionadas milhões de páginas *web* aos motores de pesquisa e milhões de utilizadores interagem com elas todos os dias. Todo o histórico comportamental da navegação do utilizador é registado nos ficheiros de registo de acessos, os ficheiros de *log* de erros, e o conteúdo das páginas *web*, e os dados existentes nos armazéns de dados *toda*, esta informação torna-se uma fonte de conhecimento da qual as empresa tentam tirar partido através das técnicas de *Web Mining*.



Até ao momento, como foi descrito, as técnicas de *web mining* dividem-se em três campo principais de acção: *Web Content Mining*, que permite extrair conhecimento do conteúdo das páginas *web* (textos, gráficos, imagens), o *Web Structure Mining*, que permite extrair conhecimento da organização de *links* da *Web*, e *Web Usage Mining*, que permite extrair padrões interessantes dos logs dos servidores *web*. Dentro deste três campos existem inumeras aplicações das suas técnicas que vão desde do comércio electrónico, a pesquisas de informação nos motores de pesquisa.

Paralelamente ao desenvolvimento das técnicas de *web mining*, foram desenvolvidas outras tecnologias nomeadamente o XML e do RDF, que vieram contribuir para para o avanço da pesquisa de informação e para a recolha de dados. Estas duas tecnologias estão na base do desenvolvimento de outros conceitos como *Semantic Web*, que tem como objectivo compreender a informação existente na *Web*. Este conceito é neste momento o centro da maior parte da investigação que se faz na *Web*, e acredita-se que, a curto prazo, passaremos do conceito de comércio electrónico a negócio electrónico (com os *sites* a negociarem directamente entre si, sem intervenção do utilizador).

No decorrer deste trabalho foram encontrados diversos obstáculos à sua realização, pelo facto de ser uma área relativamente recente e de não temos acesso à maior parte das ferramentas, o que torna o estudo mais empírico e menos prático. Também muita da informação encontrada estava desatualizada e por vezes contraditória, dado que o assunto referente a este trabalho é um campo de investigação que está em constante renovação. Como as empresas ligadas a esta área o que pretendem é vender os seus

produtos, tendem a oferecer as suas soluções como soluções mágicas, esquecendo-se que por trás das técnicas de *data mining* existe todo um processo de apoio e que depois de aplicadas há que as integrar na empresa analisando os seus resultados.

Apesar das dificuldades, foi feita uma análise crítica profunda à informação, especialmente oriunda da *Web*, tentando extrair a informação importante e útil, tendo muitas vezes que limpar informação incorrecta e completar outra inexistente. Poder-se-ia afirmar que esta foi a parte prática, uma vez que se realizou um pouco de “*web mining* manual”.

O futuro irá passar cada vez mais pela *Internet* e com a velocidade das inovações tecnológicas, as técnicas que permitem prever um futuro próximo serão cada vez mais necessárias às empresas, para que estas possam atrair e manter clientes e para se manterem a frente da sua concorrência. No entanto o futuro não se resume a isto, existem outras áreas de investigação que estão correntemente a serem investigadas desde da representação de dados, a pesquisas complexas, ao modo de uso de informação complementar, à utilização de dados estatísticos. Conjuntamente com estas áreas existem outros desafios ao processo de conhecimento como: a dimensionalidade e a sobreposição dos dados, o seu significado estatístico, a mudança constante dos dados e do conhecimento, os dados que faltam ou que estão incompletos, o ruído existente nos dados, as complexas relações entre as variáveis, o pré-processamento dos dados, a compreensão e avaliação dos padrões descobertos, a interacção homem máquina, as questões de segurança e a integração com outros sistemas. Sendo assim este campo de investigação está constantemente a ser alterada e estão sempre a surgir idéias novas, sendo por isso o seu futuro imprevisível.

Referências

✚ [DWINFOCENTER, 2000] Equipe - "The Data Warehousing Information Center - *Data Mining*", <http://www.dwinfocenter.org/datamine.html>.

✚ [THEARLING, 2000] THEARLING, Kurt - "*Data Mining, CRM, Decision Support, and Database Marketing (Kurt Thearling)*", <http://www3.shore.net/~kht/text/dmwhite/dmwhite.htm>

✚ [GIL&RODIGUES,2001] Gil, Nuno e Rodrigues, Ricardo "Text Data Mining – Seminário de Análise Inteligente de Dados" www.dei.uc.pt/backoffice/files/1008772001.ppt

✚ [Advisor, 2002] "Data Mining for E-Business" <http://www.advisor.com/Articles.nsf/aid/FRASS252>

✚ [Silva,2000] Silva, Ediberto Magalhães - "Avaliação do estado da arte e produtos *Data Mining*" [http://www.mestradoinf.ucb.br/aluno/esilva/Avaliação do estado da arte.html](http://www.mestradoinf.ucb.br/aluno/esilva/Avaliação%20do%20estado%20da%20arte.html)

✚ [Afonso,2001] Afonso, Margarida Maria Ramos, "Semantic Web", Junho 2001

✚ [Borges,2000] Borges, José Luis Cabral de Moura, "*A Data Mining Model to Capture User Web Navigation Patterns*", Julho 2000

✚ [Ramos,1998] Ramos, Carlos, "Introdução à Inteligência Artificial e aos Sistemas Baseados em Conhecimento", 1998

✚ [Spiliopoulou, 99] Spiliopoulou, Myra "*Data mining for the Web*", 1999

✚ [Fayyad et al., 1996] , Usama M. Fayyad, Gregory Piatelsky-Shapiro, Padharaic Smyth, Ramasamy Uthurusamy, "*Advances in Knowledge Discovery and Data mining*"

Bibliografia

Links visitados:

<http://www.kdnuggets.com>

<http://www.w3c.org>

<http://www.dmq.org/>

<http://www.dataspaceweb.net/>

<http://www.acm.org/sigkdd/>

<http://www.datamine.co.nz/dm.htm>

<http://www.accrue.com/index.html>

<http://www.angoss.com/>

<http://www.bluemartini.com/index.jsp>

<http://www.spss.com>

http://www.quadstone.com/info/press/2000/01_25.xml

<http://www.datamining.com/dmsuite.htm>

<http://www.lumio.com/products/?sID=13d46b38c1ba89f7f926a360ba76a32f>

<http://megaputer.com>

<http://www.microstrategy.com/Solutions/Applications/WTAM/>

<http://www.netgen.com/index.cfm?section=solutions&file=emetrics>

<http://www.netgen.com/>

<http://www.ecomminer.com/index.html>

<http://www.sas.com/products/webhound/index.html>

<http://webtrends.com>

<http://xore.com>

<http://metadatamining.com>

<http://www-3.ibm.com/software/webserver/>

<http://www.cs.ualberta.ca/~tszhu/webmining/webmyreading.htm>

http://www-personal.umich.edu/~wfan/text_mining.html

<http://maya.cs.depaul.edu/~classes/ect584/lecture.html>

<http://citeseer.nj.nec.com/231213.html>

<http://www.sims.berkeley.edu/~hearst/talks/data-mining-panel/index.htm>

<http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter9/sld001.htm>

<http://www.computerman.com.br/artigos/art21.htm>

<http://www.andersen.com/website.nsf/content/EuropePortugaleBusiness2?OpenDocument>

<http://www.sondabrasil.com.br/internet/>

<http://www.estudar.org/>

<http://www.dei.uc.pt/lei/aid/index-body.php?cadeira=54&seccao=projecto>

<http://www.nri-ltd.com/pagerank.asp>

<http://hci.stanford.edu/~page/papers/pagerank/>

www.goodlookingcooking.co.uk/PageRank.pdf

http://www.google.com.br/intl/pt/why_use.html

<http://www.inf.ufrgs.br/~wives/portugues/textmining.html>

http://www-personal.umich.edu/~wfan/text_mining.html

<http://www.sims.berkeley.edu/~heerst/talks/data-mining-panel/index.htm>

<http://www.cs.ualberta.ca/~zaiane/courses/cmp690/slides/Chapter9/sld001.htm>

Livros:

- ✚ *Data Mining Your Website*
Jesus Mena
Digital Press
- ✚ *Advances in Knowledge Discovery and Data mining*
Usama M. Fayyad
Gregory Piatelsky-Shapiro
Padharaic Smyth
Ramasamy Uthurusamy
- ✚ *Programação na World Wide Web com CGI's*
João Garrott
António Ferreira
FCA - Editora de informática
- ✚ *Tecnologia de Bases de Dados*
José Luís Pereira
FCA - Editora de informática
- ✚ *Programação Web com Active Server Pages*
João Vieira
Edições Centro Atlântico
- ✚ *Web Usage Mining for Web Site Evaluation*
Myra Spiliopoulou
Agosto de 2000
- ✚ *Automatic personalization based on Web Usage Mining*
Bamshad Mobasher
Robert Cooley
Jaideep Srivastava
Agosto de 2000
- ✚ *Prospecção dos Dados de Acesso ao Público On-line*
Paulo Batista
Mário J. Silva
Agosto de 2000
- ✚ *XML Miner, XML Rule and Metarule*
Andrew N Edmonds
Maio de 2001

- Drinking from the Firehose
Robert Cooley
- Apontamentos da disciplina de Sistemas Inteligentes
- Apontamentos da disciplina de Inteligência Artificial
- Apontamentos da disciplina de Sistemas Periciais
- Apontamentos da disciplina de Agentes Inteligentes e sistemas Cooperativos

Anexos

Produtos Data Mining

Existem inúmeras ferramentas para *Data Mining*, bem como diversos fornecedores e fabricantes. As principais ferramentas e respectivos fabricantes são listados abaixo.

Aira	GoDigital	55 051 982-5124 (Brazil)
ALICE d'Isoft	Alice	33 1 69 35 37 37 (France)
AnswerTree	SPSS	(800) 543-2185
AT Sigma Data Chopper	Advanced Technologies	(805) 872-4807
Athena Knowledge Server	Triada	(313) 663-8622
Atlas	Sofresud	33 4 94 11 57 10 (France)
BusinessMiner	Business Objects, Inc.	(800) 705-1515
Capri	MineIT Software Ltd	44 0 1232 368875 (UK)
CCM Data Correlation Model	Applied Technical Systems	(360) 478-2710
C5.0	RuleQuest Research	61 2 9449 6020 (Australia)
Clementine	SPSS	(800) 543-2185
Clustan	Clustan Ltd.	44 131 337 1448 (UK)
D-Miner	Dialogis	49 2241 206421 (Germany)
Data Mining Components	Numerical Algorithms Group	(630) 971-2337
DataBase Mining Marksman	HNC Software Inc.	(619) 546-8877
DataDetective	Sentient Machine Research	31 20 6186927 (Holland)
DataEngine	MIT GmbH	49 2408 94580 (Germany)
DataLogic/R	REDUCT & Lobbe Technologies Inc.	(306) 586-9408
DataMite	Logic Programming Associates	0181 871 2016 (UK)
DataX	Zaptron	(650) 966-8700
DBMiner	DBMiner Technology	(604) 291-537
Decision Force	Pro-Action	30 1 5254518 (Greece)
Decision List Learner	Schenley Park Research	(412) 687-4478
Decision Series	Accrue Software	(510) 580-4500
Dowser	Redshed Software	
Enterprise Miner	SAS	(919) 677-8000
Galvano	PMSI	(33 1) 45 35 87 99 (France)
Genio Miner	Hummingbird Communications	(416) 496-2200
Heatseeker	WhiteCross Systems	(310) 577-8188
iData Analyzer	Information Acumen	(612) 310-5246
Intellix Analyzer	Intellix	45 70 23 37 00 (Denmark)
Intelligent Miner	IBM	(800) 426-2255
iUnderstand	BioComp Systems	(800) 716-6770
Kingsington	InferSence	44 (0) 20 7504 6917 (UK)



Knowledge Access Suite	<u>Information Discovery</u>	(310) 937-3600
KnowledgeMiner	<u>Script Software</u>	(530) 546-9005
knowledgeSTUDIO	<u>ANGOSS Software International</u>	(416) 593-1122
kTree Knowledge Suite	<u>Kbase</u>	(888) 706-0565
KXEN Components	<u>KXEN</u>	(650) 712-1170
MineLogic	<u>Crusader Systems</u>	27 (0) 21 880 1677 (Africa)
ModelMAX Plus	<u>Advanced Software Application</u>	(412) 429-1003
ModelQuest Enterprise	<u>AbTech Corporation</u>	(804) 977-0686
Nuggets	<u>Data Mining Technologies, Inc.</u>	(516) 692-4500
ODBCMINE	<u>Intelligent Systems Research</u>	(773) 989-0426
Omega	<u>KiO</u>	
Oracle Data Mining Suite	<u>Oracle</u>	(800) 672-2531
ORESME	<u>The Galileo Company</u>	(248) 258-9657
Pattern	<u>Magnify, Inc.</u>	(708) 383-7002
PolyAnalyst	<u>Megaputer Intelligence Ltd.</u>	(812) 325-3026
prudsys Discoverer	<u>Prudential Systems Software GmbH</u>	49 (0)37153 47 1 23 (Germany)
Rough Set Data Mining System	<u>Sand Technology</u>	(514) 624-1324
PV/FutureView	<u>Continuum Software, Inc.</u>	(617) 932-8400
S-Plus	<u>Insightful</u>	(800) 569-0123
Scenario	<u>Cognos</u>	(800) 426-4667
SmartMiner	<u>GRIMMER Logiciels</u>	33 (0)1 53 09 27 50 (France)
SphinxVision	<u>ASOC AG</u>	49 0781- 96 92 96-0 (Germany)
SuperQuery	<u>Azmy Thinkware Inc.</u>	(201) 947-1881
Syllogic Data Mining Tool	<u>Syllogic</u>	310306354888(Netherlands)
TeraMiner Stats	<u>NCR</u>	(937) 445-5000
ThinkBase	<u>Science in Finance</u>	(44) 1908-584130 (UK)
VisiRex	<u>CorMac Technologies</u>	
watson	<u>Xanalys</u>	44 1625 418950 (UK)
WizWhy	<u>WizSoft, Inc.</u>	(508) 620-4554
Xaffinity	<u>Exclusive Ore</u>	(215) 643-3110
Xeno	<u>Infocentricity</u>	(415) 209-0517
XML Miner	<u>Scientio</u>	(44) 1908-584226 (UK)
Xpert Rule Analyser, Profiler	<u>Attar Software</u>	(508) 456-3946
Zoom 'n View	<u>SkyGate Development ApS</u>	45 33 339110 (Denmark)

URL: <http://www.dwinfocenter.org/datamine.html> - Atualizada em 04/2/02

Análise da Web

Esta ferramentas oferecem capacidade analíticas para além da usual análise aos logs, por exemplo análise de *clickstreams*. As ferramentas fazem um análise de tipo *Web mining*, *data webhousing*, *e-Business intelligence*, *e-Business analysis*, e *e-intelligence*.

<http://www.dwinfocenter.org/ecommerce.html> - atualizado 25/3/02

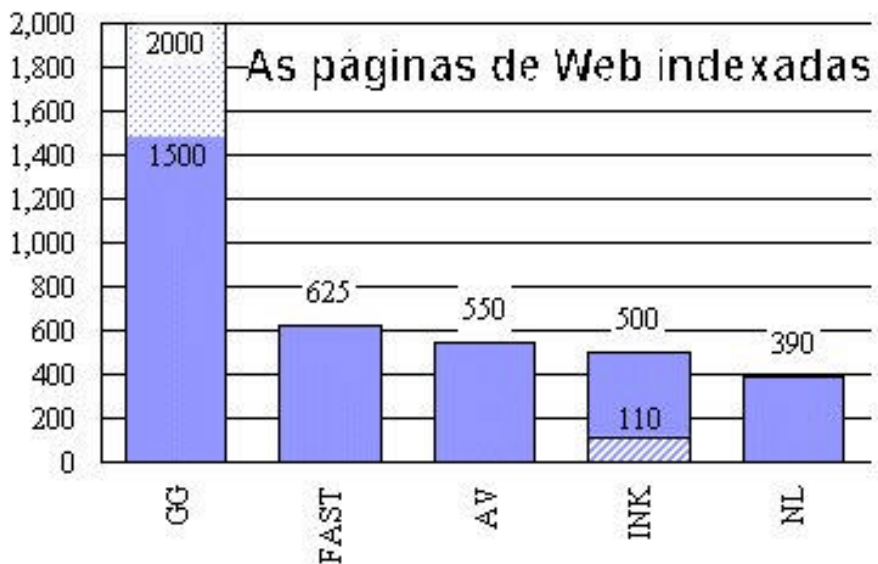
Accrue Insight	Accrue Software	(510) 580-4500
Arc 360°	Fulcrum Analytics	(888) 421-6655
Aria Enterprise	Macromedia	(415) 615-6400
Buystream Merchant	Buystream.com	(800) 261-1726
C-Insight	MetaEdge	(408) 752-9977
Centrport Advance	Centrport	(203) 341-8000
Clickstream Absolute	Clickstream Technologies	44 1223 233799 (UK)
Commerce Intelligence	InterWorld	(877) 326-6637
Customer Knowledge Platform	comScore Networks	(703) 438-2050
CustomerConversion	Quadstone	(617) 753-7393
Delano Customer Discovery	Delano Technology	(905) 947-2222
digiMine	digiMine	(425) 896-1700
E-Commerce	Broadbase	(650) 614-8300
E-Commerce Reporting & Analysis	E.piphany	(650) 356-3800
e.Intelligence	e.Intelligence	(952) 920-0478
EasyMinerWeb	MINEit Software Limited	44(0) 28 90368875 (UK)
eBizinsights	Visual Insights	(630) 753-8600
eChannel Advisor	Active Decisions	(650) 342-0500
eConsumer	Innovative Resource Group	(412) 781-7400
ECRM	Informatica	(800) 653-387
eLuminate	Coremetrics	(877) 721-2673
elytics.com Analysis Suite	elytics.com	(617) 492-7760
EMine	SPSS	(312) 651-3000
Enterprise ChannelMetrics	OneChannel	(650) 404-0120
Enviz	Enviz	(650) 298-0900
eSpective	Telemate.net	(770) 936-3700
Essentials	Personify	(415) 782-2050
Genalytcs	Genalytcs	(978) 465-6373
Hitbox Enterprise	WebSideStory	(858) 546-0040
HitsIntoLeads	Netmining	32 16 38 72 50 (Belgium)
Hyperion e-Marketing Analysis	Hyperion	(408) 744-9500
iLux Enterprise	iLux	(510) 226-5600
InfoSplit Market Reports	InfoSplit	(212) 683-0662
Insight Reporting	Primary Knowledge	(212) 233-2110
IntelliLog	LogMetrix	33 1 44 88 93 88 (France)

Intellitracker	<u>Intellitracker</u>	44 020 7665 1624 (UK)
Lateral Line	<u>Appliant</u>	(877) 227-7542
LimeLight	<u>Key Lime Software</u>	(858) 509-0055
LiveStats	<u>MediaHouse Software</u>	(819) 776-0707
MicroStrategy Web Business Analyzer	<u>MicroStrategy</u>	(800) 848-8600
Mpi	<u>veridiem</u>	(978) 461-2444
Narus Intelligence	<u>Narus</u>	(617) 665-9200
net.Analysis	<u>net.Genesis</u>	(617) 665-9200
NetAuditNow	<u>NetAuditNow</u>	(609) 683-3800
NetTracker ebusiness Edition	<u>Sane Solutions</u>	(800) 407-3570
Oracle Clickstream Intelligence	<u>Oracle</u>	(800) 633-1071
Pivotal Digital Intelligence	<u>Pivotal</u>	(877) 748-6825
Plexus	<u>Hanrick Associates</u>	(206) 381-0602
prudsys ECOMINER	<u>Prudential Systems Software GmbH</u>	49 (0) 3 715347123 (Germany)
RedSheriff Measurement	<u>RedSheriff</u>	(212) 297-6221
RedTrack	<u>Red Eye International</u>	44 20 7627 9300 (UK)
RTMetrics	<u>AuriQ Systems</u>	(626) 564-2781
SageAnalyst	<u>SageMetrics</u>	(866) 723-9473
SiteKeyZ	<u>Complex Systems</u>	01 42 21 40 80 (France)
SiteStat	<u>NedStat</u>	(212) 625-9700
Stratum e-Business Performance Management	<u>Silvon Software, Inc.</u>	(800) 874-5866
SuperStats Enterprise	<u>MyComputer.Com</u>	(877) 722-7055
SurfAid	<u>IBM</u>	(800) 426-2255
Synera ePack	<u>synera</u>	(952) 814-9300
Urchin 3	<u>Quantified Systems</u>	(619) 233-1400
WatchWise	<u>WatchWise</u>	(858) 509-0055
Web Hound, Web Mining, Intellivisor	<u>SAS Institute</u>	(919) 677-8000
Web Utilization Miner	<u>Humboldt University</u>	
WebAbacus	<u>WebAbacus</u>	44 (0)20 7430 1515 (UK)
webfeedback	<u>Ingenieurbüro Liebhart</u>	61 3 93482441 (Switzerland)
webHancer e-Business View	<u>webHancer</u>	(617) 621-4075
WebHouse	<u>Torrent Systems</u>	(617) 354-8484
WebInsight	<u>Applix</u>	(508) 870-0300
Weboscope	<u>Weborama</u>	01 53 01 49 20 (France)
WebSuxess	<u>Exody E-Business Intelligence</u>	49 6196 9599 0 (Germany)
WebtraffIQ	<u>WebtraffIQ</u>	44 (0) 789 999 2288 (UK)
Webtrends Commerce Trends	<u>Webtrends</u>	(503) 294-7025
WebVision - Marketing	<u>NetAcumen</u>	(650) 696-3100
WX/Web Analytics	<u>White Cross</u>	(415) 908-6883
Xelector Web Activity Analysis	<u>Xelector</u>	353 (1) 648 6200 (Ireland)

Motores de Pesquisa

É muito importante que o site seja adicionado a um motor de pesquisa. Com a explosão da *Internet*, o *marketing* de um *site* pode fazer a diferença entre um óptimo *site* que ninguém vê e um *site* que cativa os clientes, e que pode ser facilmente encontrado, obtendo uma boa classificação nos motores de pesquisa. A empresa deve receber relatórios que incluam a classificação do motor de pesquisa, as estatísticas mensais, o número de páginas que são mais vistas pelos utilizadores.

No seguinte gráfico estão representadas as milhões de páginas que são adicionadas à *Internet* todos os dias, o *marketing Internet* passa a ter um papel muito importante nos negócios na *Web*. A escolha de motor de pesquisa pode ser igualmente importante.



Legenda:

GG=Google, FAST=FAST,
AV=AltaVista, INK=Inktomi,
NL=Northern Light.

Percentagens divulgadas pelos motores de
pesquisa em 11 de dezembro de 2001.
www.SearchEngineWatch.com

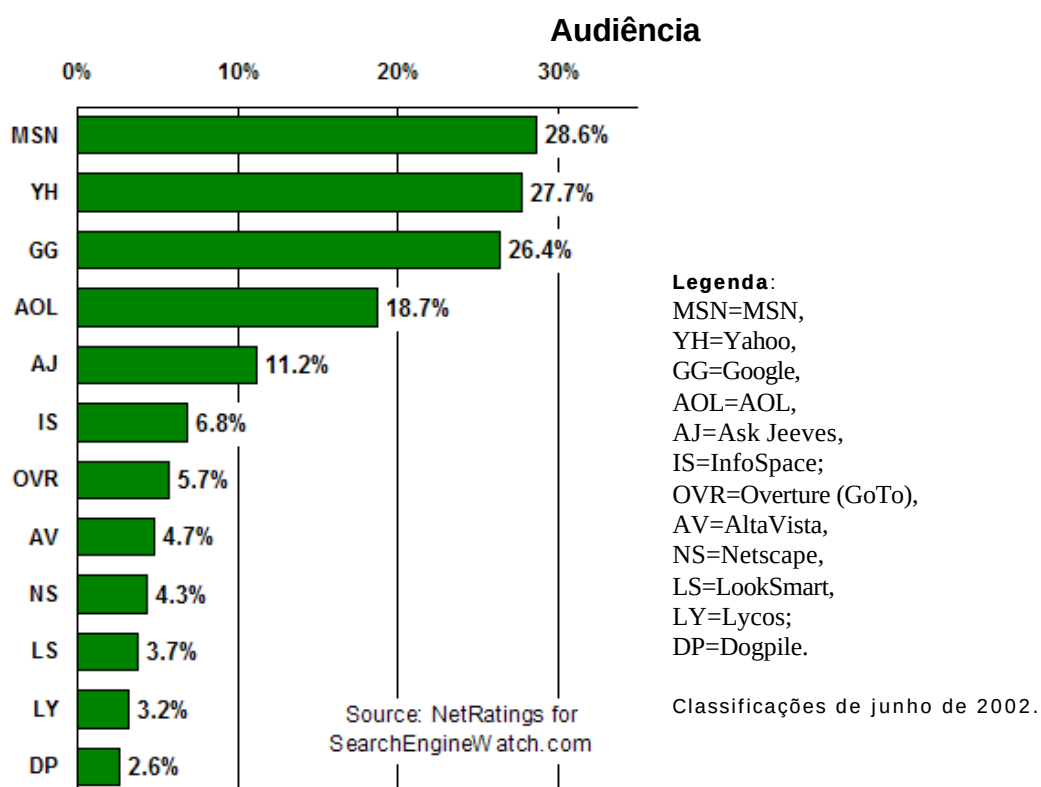
Algumas das pesquisas efectuadas pelos motores de pesquisa são feitas através de palavras chave, a que os produtos ou serviços dos sites estão relacionadas. Se a página do *site* não aparece no *top* do 10 ou 30 primeiros e a concorrência aparece, não importa o número de motores de pesquisa em que esteja inscrito, a concorrência está a ganhar.

Uma das palavras que faz com que a página obtenha uma boa classificação no motor de pesquisa é a palavra “grátis”, atraindo assim tráfego para o site.

Métodos de promoção do Web Site	
Classificação de motor de pesquisa	66%
E-mail	54%
Brochuras impressas	42%
Catálogos impressos	40%
Feiras	37%

Tabela IV: Metodos de promoção do Web Site

O seguinte gráfico mostra o número de pessoas que visitaram os motores de pesquisa durante um mês. Algumas pessoas podem ter visitado mais do que um motor de pesquisa, o que explica que o total das percentagens exceda os 100%.



 all the web, all the time http://www.alltheweb.com	AllTheWeb.com (também conhecido por FAST Search) é um dos maior indexes da <i>web</i>. FAST oferece indexes de multimedia e indexes <i>mobile/wireless</i>. Os resultados do FAST são dados a inúmeros portais, incluindo Terra Lycos. FAST Search foi lançado em maio de 1999.
 http://www.yahoo.com	Yahoo é o motor de pesquisa mais popular da <i>web</i> para encontrar facilmente informação. Yahoo tem mais de um milhão de sites inscritos. Yahoo também mostra resultados do Google. Se a pesquisa falha, o Yahoo compara os resultados do Google e estes são mostrados. Os resultados do Google são mostrados depois das do Yahoo. Yahoo é o motor de pesquisa mais antigo tendo sido lançado em 1994.
 http://search.msn.com/	O motor de pesquisa MSN da Microsoft é uma compilação do LookSmart Directory e AltaVista. Os dados do RealNames e Direct Hit também estão disponíveis. MSN permite aos utilizadores do Internet Explorer 5 ou superior de guardar pesquisas anteriores.
 http://www.aol.com/	O motor de pesquisa AOL permite pesquisar na <i>Web</i> e na lista de conteúdos do AOL. A lista principal de categories e sites vem do <i>Open Directory</i>. Google também dá resultados, como <i>backup</i> a informação do <i>directory</i>. A colocação paga de links do <i>GoTo</i> também está disponível no motor de pesquisa AOL. Antes de ser como <i>AOL Search</i> em Outubro de 1999, o <i>AOL search</i> era Excite-powered AOL NetFind.
 http://www.lycos.com/	Lycos começou como um motor de pesquisa, dependendo das listas da <i>Web</i>. Em Abril de 1999, passou a ter um modelo de directório semelhante ao do Yahoo. As suas principais listas vinham do Open Directory project, e as secundárias dos resultados do Fast/All da Web. Em Outubro de 1998, Lycos comprou o motor de pesquisa HotBot, que utiliza Inktomi.
 http://home.netscape.com/	Os resultados do motor de pesquisa da Netscape vêm principalmente do Open Directory e da base de dados Smart Browsing da Netscape, que contem os site oficiais. Os resultados secundários vêm do Google.

 http://www.google.com/	O Google é o motor de pesquisa mais popular da actualidade, ele pesquisa que utiliza a popularidade dos <i>links</i> para classificar os <i>sites</i>. Isto pode ser muito útil para encontrar <i>sites</i> em pesquisas genéricas, pois os utilizadores na <i>Web</i> votam nesse <i>sites</i> colocando <i>links</i> para eles. Google mostra também resultados do Yahoo. O Google permite pesquisar imagens através do Usenet e de uma versão própria do Open Directory.
 http://www.excite.com/	Excite tem o seu próprio index, e também utiliza o do LookSmart's Directory para resultado baseados em categorias. Excite foi lançado em 1995.
 http://www.iwon.com/	iWon tem por base Inktomi, DirectHit, LookSmart e RealNames. iWon também mostra os links pagos do Goto.com.
 http://www.nbc.com/	NBCi tem um directório de <i>web sites</i>, e também mostra resultados do Inktomi. Mostra os links pagos do Goto.com.
 http://www.overture.com/	GoTo vende espaço para anúncios com os resultados das pesquisas. A publicidade é paga por click. O top 2 são listados no AltaVista, AOL Search, Direct Hit, GoTo, HotBot, iWon, Lycos, NBCi, Netscape Search, Search.com. Mais tarde foi chamado de Overture.
 http://www.askjeeves.com/	Ask Jeeves leva o utilizador à página exacta que corresponde a pesquisa. Se a pesquisa falha, o motor de pesquisa vai buscar resultados a outros motores de pesquisas. Direct Hit é a base das pesquisas por categoria. Ask Jeeves também utiliza GoTo.com.
 http://www.altavista.com/	AltaVista é um dos maiores motores de pesquisa na web, em termos de páginas indexadas. Também tem inúmeras características para atrair utilizadores, o serviço de directórios é baseado no LookSmart Directory. AltaVista também mostra os links do GoTo. AltaVista foi lançado em Dezembro de 1995.
 http://www.looksmart.com/	LookSmart contém um directório de web sites. Para além de ser um serviço por si só, dá resultados a outros motores de pesquisa como MSN Search, Excite, etc. AltaVista fornece resultados ao LookSmart quando a sua pesquisa falha. LookSmart foi lançado em Outubro de 1996.

 http://www.teoma.com	Direct Hit é uma empresa que trabalha com motores de pesquisa de forma a refinar as suas pesquisas. Monitoriza os sites que são acedidos pelos utilizadores quando estes estão a ver os resultados da pesquisa. Os sites mais acedidos são classificados pelo Direct Hit. Recentemente o Direct Hit passou a ser chamado de Teoma.
 http://www.dogpile.com/	Dogpile compila resultados de pesquisa de diversos motores de pesquisa incluindo GoTo.com, About Sprinks, LookSmart, DirectHit, etc.
 http://www.hotbot.com	O resultados do HotBot vem do Direct Hit, e os resultados secundáriosvem do Inktomi. A sua informação vem do Open Directory. HotBot foi lançado em Maio de 1996.
 http://www.inktomi.com	O indice Inktomi foi criado para alimentar o HotBot. Agora o motor de pesquisa Inktomi também alimenta outros motores de pesquisa. Não se pode pesquisar directamente no Inktomi só através dos seus associados.
 http://dmoz.org/	O Open Directory utiliza editores voluntários para catalogar a Web. Tendo sido lançado com NewHoo em junho de 1998.

Ferramentas Adicionais

Accrue HitList



<http://www.accrue.com/index.html>

É uma poderosa e flexível ferramenta de análise de de ficheiros de registo de acessos com mais de 300 relatórios.

Accrue HitList é uma solução de análise, online e *offline*, de dados em tempo real. HitList otimiza o efeito das iniciativas na *Web* fazendo decisões de *merchandising* baseadas em dados em dados armazenados implicando o aumento do lucro e a satisfação do cliente.

Para fazer com que as decisões de investimento na *Web* em negócios fundamentais as empresas tem que medir o impacto do *merchandising* e da publicidade nessas decisões. As empresas precisam de capturar a informação do tráfego no website e com os dados demográficos otimizar os seus esforços de *marketing* e *merchandising*.



FIGURA 26: ACCRUE HITLIST

Serviços Educacionais da Accrue

Accrue fornece uma gama variadas de técnicas de treino:

- ⚡ *Web-based*
- ⚡ *Classroom-lead*
- ⚡ *On-site*
- ⚡ Técnicas e dicas

Serviços Profissionais da Accrue

A Accrue desenvolveu serviços de suporte para o cliente como *setup*, treino, e manutenção de sistema:

■ Instalação

⚡ Serviço de instalação rápida

Permite que o cliente instale e configure correctamente os produtos da Accrue assegurando que o sistema carrega para a base de dados toda a informação bem como a demonstração como correr os relatórios

Serviços de Implementação

A instalação de serviços começa com uma análise completa e instalação dos relatórios dos produtos, plano de trabalho, criação do grupo da URL, filtros, configuração de *setup*, administração de sistema e análise de negócio. O resultado é uma instalação configurada e personalizada.

Serviços de Manutenção preventiva

QuickTune

QuickTune inclui um *upgrade* da corrente instalação do Accuer Insight ou Hit List devido a mudança do volume de tráfego ou a forma como os dados são capturados e reportados.

QuickCollector

QuickCollector inclui instalação e/ou configuração de uma única Accrue Insight ou Hit List Network Collector.

QuickWarehouse

A *QuickWarehouse* inclui Instalação e/ou configuração de uma única Accrue Insight e Hit List Data Warehouse.

QuickUpgrade

A *QuickUpgrade* inclui um *update* da ultima versão do Accrue Insight e Hit List. Este serviço inclui Instalação e/ou configuração de uma única do Accrue Network Collector e Data Warehouse.

Solução personalizada

Personalização de Relatórios

Suporta a personalização de relatório usando interfaces standard e processamento em *batch* e distribuição automática. Oferece aos utilizadores uma análise mais detalhada.

Integração dos dados

Os produtos Accrue oferecem uma análise completa de todos os dados dos clientes da *Web*. Os utilizadores podem integrar dados externos na base de dados de clientes e outros tipo de fonte de informação de clientes para que os programas possam ter uma visão geral dos seus negócios na *Web*

Serviço de Merchandising

Oferece análise e recomendações de *Web merchandising* baseados na informação dos clientes compilados pelos produtos Accrue. Este serviço ajuda os clientes a aprender como ganhar previsões no seu ambientes específico de negócio e obter resultados que mapeiam a estratégia de negócio.

Conteúdo das Bridges

O conteúdo das Accrue *Bridges* asseguram que Accrue Insight e *Insight* possam adicionar módulos que fornecem uma

compreensiva análise de *e-Business*, para uma largo leque de tecnologias e abordagens que entregam conteúdos e serviços na *Web*. Para tirar o maior partido de cada produto, o programa de *bridges* usam uma mistura apropriada de produtos e serviços. Por exemplo, a bridge pode ser a oferta de um produto, a combinação entre de *templates* de *software* a personalizações efectuadas pelos clientes ou pelo Accrue Consulting Group.

Cookie Callout

Coloca o utilizador da *cookie* no inicio do campo da *cookie*. Este serviço é para clientes que utilizam múltiplas *cookies* e querem utilizar uma das características do Accrue Insight a monitorização de *cookie*.

Bridges Comerciais

São desenhadas para a captura de códigos de eventos e transacções dos motores de comercio e combina-os com o Accrue Insight Warehouse e relatórios de comercio que dão resultados baseados em factos .

Campaign Bridges

São desenhados para a captura de dados de publicidade dos servidores de publicidade e interagira ao dados com o Accrue Insight e os relatórios das campanhas que fazem com que o investimento seja retornado.

Web Log Analysis — Análise de tráfego

O software de análise de ficheiro de registo de acessos tem por base os ficheiro de log dos servidores *Web*. A analise procura compreender o que está a acontecer no *Web site*. Qual a quantidade de trafego que o site está a ter? Quanto pedidos falharam, e que tipo de erros estão a ser gerados?

Adicionando outro tipo de informação como as campanhas de marketing, filtros, e informação virtual do servidor. Pode responder a perguntas como:

-  Quais as empresas que visitam o *site*?
-  Quais as páginas que são mais / menos visitadas?
-  Que *sites* são originadas as visitas?

Web Mining — Combinação do trafico da Web com outras bases de dados

O *Web site* cria informação. Quando se combina os dado do trafego da Web com outras bases de dados. Relacionando toda a informação, pode-se transformar em informação util para os negócios.. Therefore, a O sistema de *Web mining* deve integrar qualquer fonte de dados correndo estas em qualquer sistema operativo.

Com a informação a mudar constantemente, o sistema deve permitir ligar a outras bases de dados, para gerar os relatórios. O *Web mining* em oposto ao sistema *Web log analysis*, tenta descobrir tendencias que de outra forma passariam despercebidas. Os relatórios são gerados com uma grande variedade de informação desde a identificação do cliente a informação do produto de uma forma que a informação possa ser compreendida.

O sistema de *Web mining* ajuda a responder as seguintes perguntas:

- Como é que a informação demográfica e psicográfica do utilizador é relacionada com o seu comportamento de navegação?
- Qual é o ROI do *Web site*?
- Quais os banners que trazem mais visitantes ao *site*?

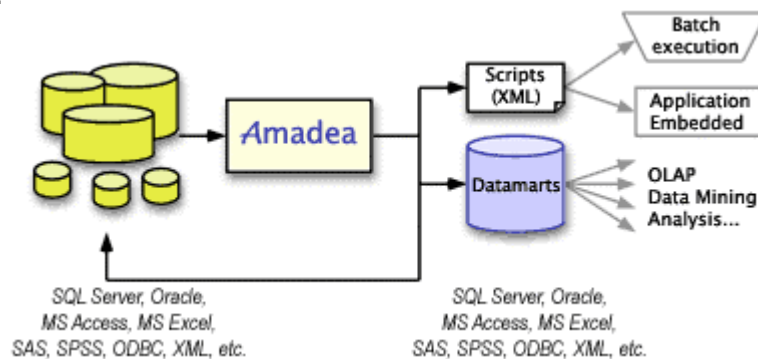
Amadea Ferramenta de Data Morphing



www.isoft.com

AMADEA, é a primeira ferramenta de *data Morphing*, e é também uma ferramenta que permite transformar os dados de diferentes proveniências e transforma-la em dados ricos em conhecimento e pronta a ser introduzida em Sistemas de negócios Inteligentes.

AMADEA permite modelar a informação de forma adaptar aos sistemas de informação e a evolução de Mercado e de estratégias. O conceito AMADEA baseai-se em cenários com parâmetros. Estes cenários são graficamente implementadas, de um modo interactivo, e sem requer programação. A preparação dos dados fica um processo interactivo e fácil de fazer manter e compreender.

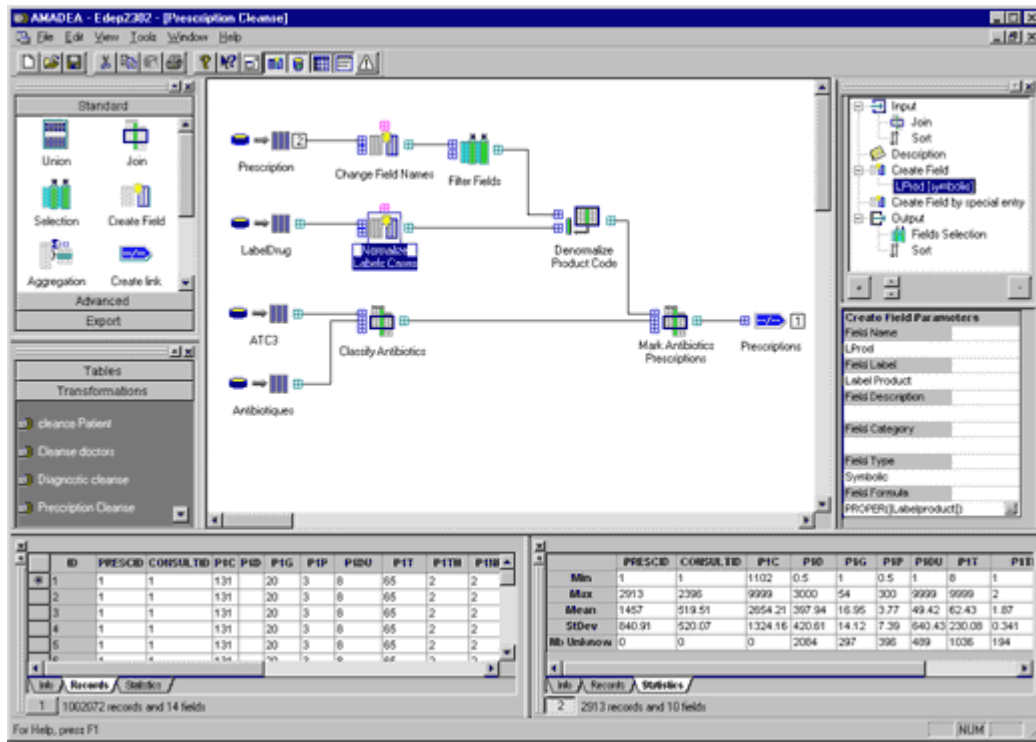


Uma vez conectado as fontes de dados, AMADEA contém funções pré-definidas para limpeza de dados (como por exemplo a substituição de dados em falta). Este passo é suportado por bibliotecas de operadores. Os operadores são graficamente unidos para que o utilizador não necessite de programar. As tabelas resultantes podem ser exportadas para os mais diferentes formatos de dados existentes no mercado.

AMADEA é baseada em *standards*, e pode ser conectada as mais usuais fontes de dados como ficheiro de acessos, ficheiros flat, RDBMS, ficheiro de XML, etc. e descreve os passos necessários para produzir e agregar dados para a análise ou definição de modelos. AMADEA pode ser utilizado em diferentes domínios como personalização de sites de *e-commerce*, estudos de comportamento de clientes de *e-commerce*, análise de vendas de retalho, CRM, etc.

Transformação como processo

A transformação dos dados já não passa por um programa complexo de computador desenhado para executar longe dos utilizadores. Sendo assim este processo é a ser simples e compreensivo.



A janela central permite ao utilizador desenhar e controlar a execução do processo de transformação. Seleccionado o *output* de um dos operadores automaticamente actualizado a grelha de informações para reflectir as informações processadas pelo operador na tabela de *input*.

Na frame da esquerda do écran dá acesso aos operadores de informação. Os parâmetros de cada operador podem ser escritos à direita do écran.

Interactividade

- Definição gráfica dos processos de transformação, não é necessária qualquer programação.
- Processo interactivo de transformação de dados através de um interface amigável.

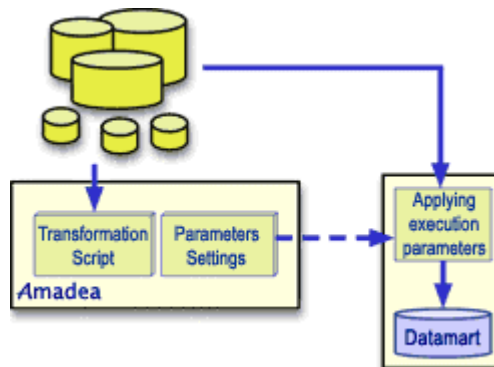
AMADEA permite:

- AMADEA permite um ganho de tempo no desenvolvimento e manutenção de complexos processos de transformação de dados

- ⚡ O Motor ETL permite a exploração interactiva do processo de transformação.
- ⚡ O uso de AMADEA aumentar a eficiência na preparação de informação até 95% do tempo normalmente necessário da forma tradicional.

Processo dinâmico com parâmetros

O processo de transformação é dinâmico e flexível. Um datamart é criado com AMADEA que posteriormente pode ser redefinido sempre que os parâmetros de que depende. Os scripts para to selecção dos parâmetros são páginas da Web que são geradas automaticamente pela definição da transformação.



Com o AMADEA é possível produzir novas variáveis, filtrar dados para que a transformação tenha apenas a informação relevante.

AMADEA consegue tratar grandes volumes de dados para isso utiliza um motor para gerir a memória e otimizar assim o uso da memória do computador de acordo com o espaço livre. O *software* adapta-se facilmente as capacidade existente podendo assim oferecer uma solução escalavel.

Características

- ⚡ Conexão a número variadas de fontes de dados: XML, ficheiro de registos de acessos, RDBMS, ficheiros flat.
- ⚡ Extracção e análise de ficheiro de registo de acessos a servidores Web
- ⚡ Possibilidade de reutilização de bibliotecas.
- ⚡ Interactividade
- ⚡ Transformação por operadores
- ⚡ Reunião de informação para o processo
- ⚡ Possibilidade de alterar os parâmetros e reutilização em processos
- ⚡ Transformação pode ser feita pelo *browser*
- ⚡ Possibilidade de seguimento do processo
- ⚡ Tratamento de grandes conjuntos de dados
- ⚡ Tratamento de metadados
- ⚡ Ganho de produtividade em 90%



<http://www.angoss.com/>

ANGOSS KnowledgeWebMiner

KnowledgeWebMiner®

KnowledgeWeb Miner é uma de solução de análise para *Web mining* e análise de dados resultantes de ambientes de e-Business.

KnowledgeWeb Miner ajuda as empresas a converter os seus dados de *Web log* em informação capaz de beneficiar das técnicas de *data mining*, sem grande investimento em *software* ou *hardware* e com baixo risco de implementação.

KnowledgeWeb Miner análise o *Web log* dando os seguintes resultados:

- O perfil do visitantes por segmento e atributos individuais
- prevê o comportamento e resultados
- aprende e responde a preferencia de um visitante
- Avalia e mede o impacto de online e offline de actividades de marketing
- acede e melhora o designe, estrutura e conteúdo

KnowledgeWebMiner utiliza os componentes de *Data mining* do *KnowledgeSTUDIO*, os algoritmos que este utiliza para análise de streams, o enriquecimento dos dados através do *Acxiom Data Network*, e o processamento online de dados geodemograficos de marketing em mais de 95% de casas nos US.

A ferramenta trabalha com ficheiro de registo de acessos do servidor *Web* e utilizando as funções de relatórios do *Web log parsing* e de outra ferramentas como: *Net*Genesis*, *WebTrends*, e *Accrue Software*. ANG OSS planeia lançar o *VisibleWeb*, um *plug-in* para *KnowledgeWebMiner*, que permite a visualização da actividade do *site* no *desktop* do utilizador.

O *KnowledgeWebMiner* corre em ambientes como *Windows NT/2000* e *SUN Solaris*. [Advisor, 2002]

Quadstone – CustomerConversion



<http://www.quadstone.com>

É uma análise centrada no cliente com análise gráfica e relatórios da *Web* e outros tipos de dados. CustomerConversion, é uma solução de *software* que permite aos *e-businesses* compreender e prever o comportamento das compras de um cliente e leva o cliente a ter o máximo lucro. CustomerConversion é um produto que analisa a *Web* e outros tipo de dados relativos ao comportamento do cliente e situação actual do mesmo - contrario a gerar relatórios estáticos de tráfego da *Web*. Compreende o comportamento do cliente e o que leva as compras *on-line*, ajustando os *e-marketer* a otimizar a aquisição de recursos, alvejando os segmentos de clientes mais lucrativos e converte os visitantes da *Web* em clientes leais e de longo prazo. O CustomerConversion, preenche o espaço entre o *e-commerce* e informação do cliente. Em vez da solução que apenas relata os *links* das páginas que foram visitados. O CustomerConversion permite as empresa a investigação com a utilização de gráficos sofisticados e interactivos, que mostram como os clientes se comportam. O estudo dos perfis de cliente em termos de comportamento, valor, lucro e preferências. CustomerConversion entrega a informação do comportamento do cliente à Quadstone. CustomerConversion permite aos vendedores entender, prever, administrar e influenciar o comportamento do cliente. Os benefícios do CustomerConversion:

- Coloca *click-streams* no indicador de comportamento para cada cliente o CustomerConversion dá uma vista do *customer-centric* de toda a informação dos *e-commerces* e do sistema de *e-marketing*, registos de transacções da *Web*, e fontes de dados de informação de clientes. A rápida e automática transformação de dados em informação de clientes que permite tomar decisões de negócios imediatas resultante a acção do cliente.
- Oferece interactividade com os clientes e visualização no *website* Quadstone e é a única solução que permite a análise da descoberto interactiva dos dados existentes usando vistas tridimensionais.
- Prevê e modela os clientes mais lucrativos, identificando os que são mais afectados pelas diferentes técnicas de *marketing*. *E-tailers* são capazes de facilmente refinar e otimizar as técnicas de *marketing* gastas e os focos principais para alcançar os clientes mais lucrativos.
- Melhora a flexibilidade, e as escalabilidade e lançamento do CustomerConversion que tem um processamento *end-to-end* transformando os dados do *click-stream* numa selecção dos clientes certos, permite aos vendedores tomar acção imediatas de *marketing*. A arquitectura paralela ajuda a escalar a análise para aumentar a velocidade e produtividade dos dados analisados.

Especificações do sistema : O CustomerConversion corre em servidores NT e Unix, podendo-se ligar a bases de dados relacionais e a outras fontes de dados.

Data Mining Suíte



<http://www.datamining.com/dmsuite.htm>

O Data Mining Suite é uma solução para as empresas e oferece o suporte a decisões em larga escala. Permite minar os dados em multi-tabelas de SQL. O Data Mining Suite trabalha directamente em repositórios de SQL sem necessidade de amostragem ou extracção de ficheiros. Acede a grandes volumes de dados nas tabelas de um servidor, consoante as descobertas cria padrões e gera automaticamente textos em inglês e gráficos bem como documentos explicativos na *Internet*.

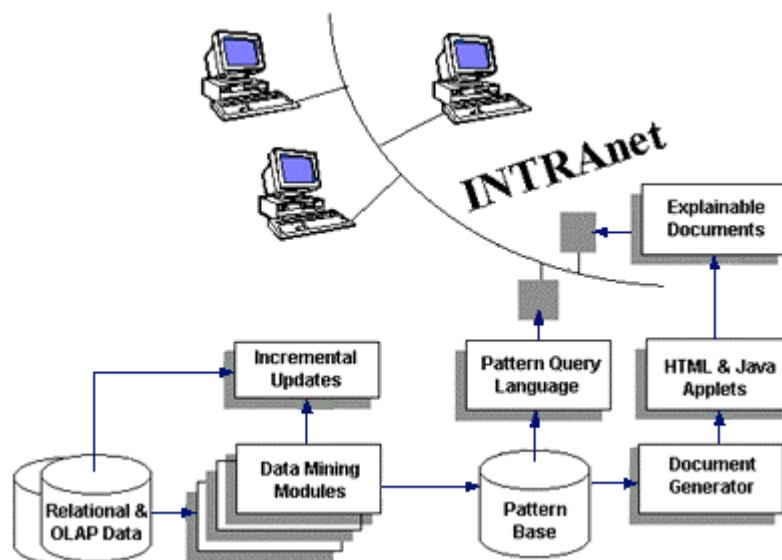


FIGURA 27: DATA MINING SUITE

A aplicação de Data Mining Suite é baseada totalmente para o suporte a decisão. A implementação do servidor é baseado na arquitectura das três camadas que permite grande escalabilidade de descoberta nas grande bases de dados de SQL com mais de 90% de desempenho no servidor.

O Data Mining Suite pode ser distinguido pelas seguintes capacidades:

- **Acesso directo a grandes bases de dados.** O Data Mining Suíte trabalha directamente com grandes bases de dados SQL, e não requer amostras nem extracções para ficheiros flat que perdem as funcionalidades do SQL o que leva a marginalização de resultados. A outra vantagem de é a capacidade de lidar com dados numéricos e não

numéricos uniformemente. O Data Mining Suíte não fixa amplitudes nos dados numéricos atecipadamente, mas descobre as suas amplitudes dinamicamente.

- **A descoberta em multi-tabelas.** O Data Mining Suíte descobre padrões em multi-tabelas em bases de dados SQL sem ter que juntar ou construir um ficheiro de extracto. Isto faz que o Data Mining Suíte tenha um papel chave na extracção de conhecimento em grandes bases de dados. Assim as varias bases de dados multitabela podem servistas juntas numa única vista.
- **Sem amostras ou extractos.** A amostragem foi inventada porque não se tinha acesso a toda a informação a ser analisada. Mas os amazens de dados oferencem este acesso. O Data Mining Suíte trabalha com toda a base de dados, mas também trabalha com amostragem se utilizador insistir.
- **Padrões.** O Data Mining Suíte descobre padrões que vão para ale das árvores de decição e simples afinidades. As árvores de decisão são muito limitadas e podem não encontrar toda a informação existente na base de dados . Sendo assim o Data Mining Suiteé baseado em regras.
- **Linguagens.** O Data Mining Suite tem uma linguagem de expressões do tipo If...then.
- **Tratamento uniforme de dados numéricos e não numéricos.** O Data Mining Suite consegue lidade com diferentes tipo de dados. Pode lidar com dado numéricos e não numericos e automaticamente descobrir amplitudes dentro dos dados numéricos.
- **Arquitectura.** O Data Mining Suite tem uma arquitectura cliente servidor em três camadas. O processamento é feito num servidor Unix.
- **Iniciativa do sistema.** O Data Mining Suite inicia o processo de *data mining* e este automaticamente foema uma hipótese baseado nos dados e converte a hipótese em expressões de SQL. O Data Mining Suite selecciona os padrões significativos e filtra os que não são importantes.
- **Transparência da descoberta e previsões.** O Data Mining Suite dá explicações como os pradões estão a ser encontrados. O mesmo acontece quando faz previsões.
- **Não é sensível ao ruído.** O Data Mining Suite não é sensível ao ruído internamente utiliza a análise de lógica *fuzzy*. Se os dados tem ruído o Data Mining Suite reduz o nivel de confiança associado aos resultados.

OK-LOG



<http://www.oklog.biz/>

- A análise é realizada a partir dos ficheiros gerados automaticamente pelo servidor Web (ficheiro de registo de acessos).
- OK-Log gera dois novos ficheiros (logplus.log e metalog.log) com a mesma estrutura que o anterior, mas contendo mais informação e organizada pelo que o cliente esta a ver no *browser*.
- OK-Log contem um ficheiro que descreve o que o utilizador esta a ver no ecrã num dado momento.
- OK-Log permite visualizar relatórios de actividade e estatísticas o que permite reconhecer as áreas do site mais atractivas para os clientes.

A figura seguinte compara as diferentes maneiras de registar a actividade do site pela maneira convencional do ficheiro de registos e depois pelos ficheiros LogPlus (logplus.log) e Log2Screen (Metalog.log).

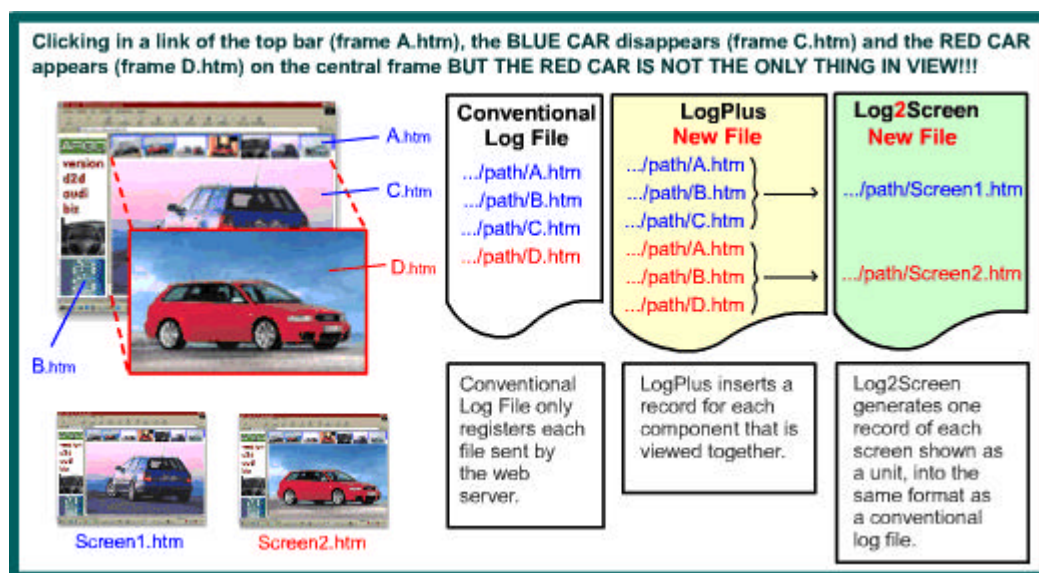
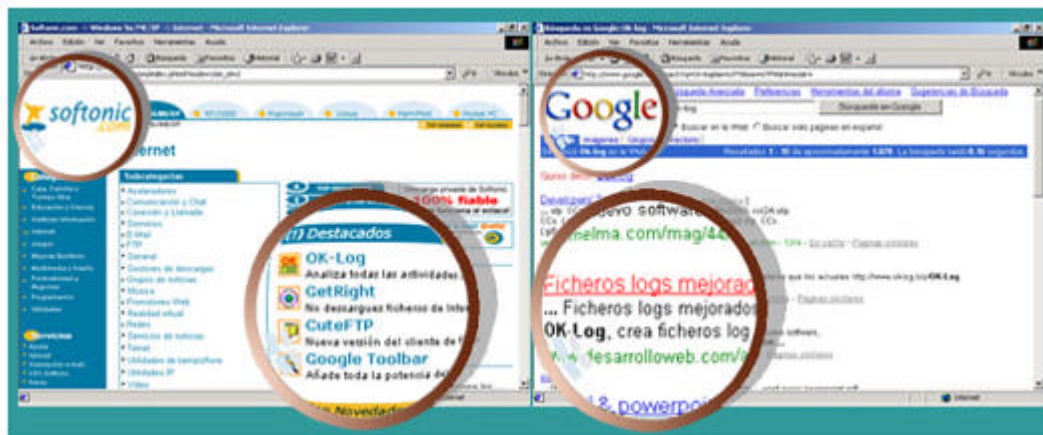


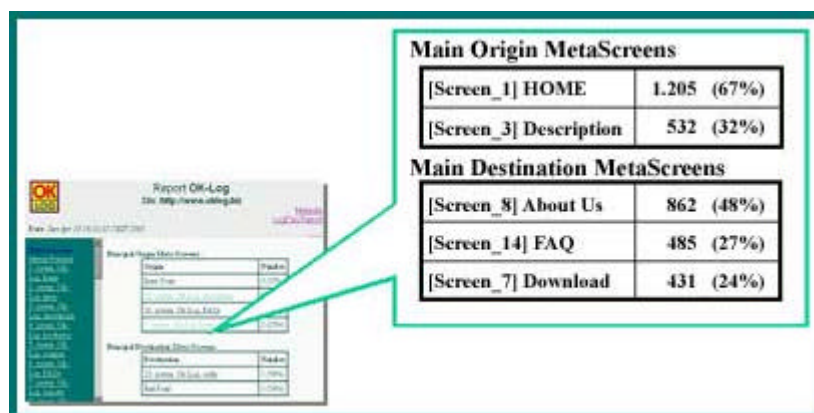
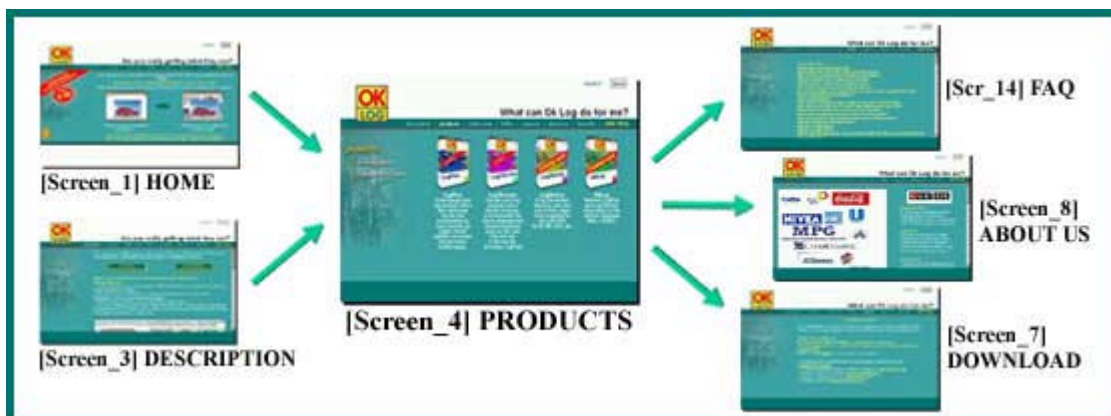
FIGURA 28: REGISTO DE ACTIVIDADE

OK-Log visita todos os *sites* externos que contem um *link* para o *site* alvo e que geraram um visita e faz uma análise exhaustiva sobre o contexto do *link* e qual a actividade principal do *site* que originou a visita.

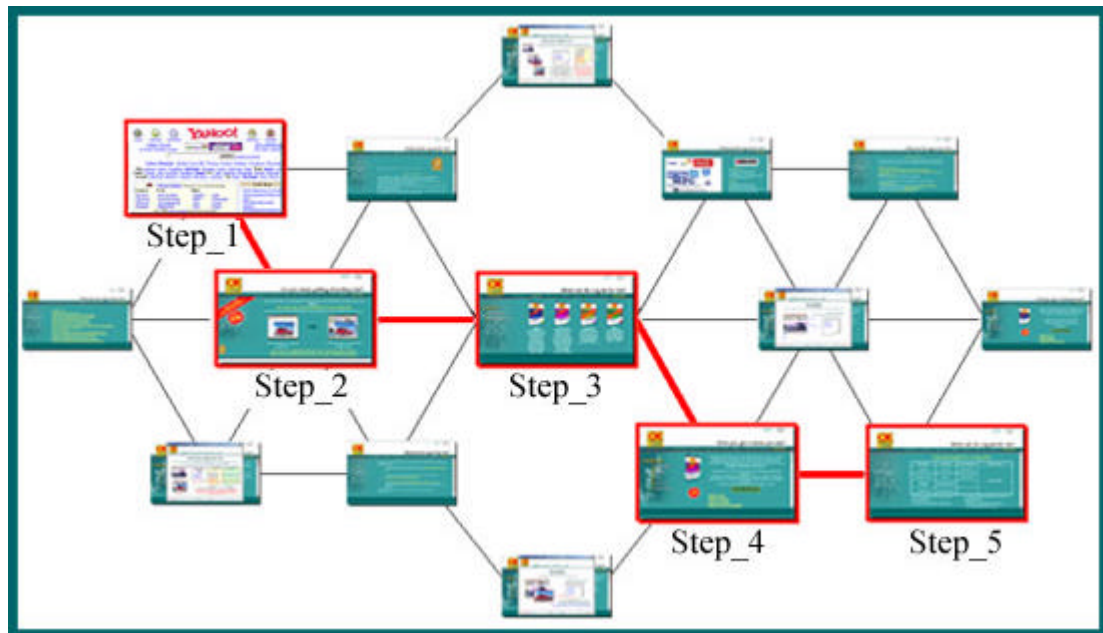


Origem e Destino

Para cada *MetaScreen*, Ok-Log gera um relatório com a origem principal e destino do *MetaScreens*.



O caminho que o utilizador utilizou



Produtos

OK-LOG v2.01 permite responder as seguintes perguntas:

- Qual a origem dos visitantes do site?
- Qual é o seu comportamento?



Benefícios

- traduz a informação do site em informação útil.
- Converte os dados técnicos em informação de marketing.
- Identificas as áreas mais visitadas do Site.

OK-Log permite a visualização do que o utilizador está a ver no écran, e dá informação adicional sobre:

- Quais são as páginas mais vistas no site
- Quais os caminhos que o utilizador utiliza para navegar no site
- Qual a media de tempo gasta em cada página
- Saber quais os links que estão em baixo antes que o utilizador chegue lá.

Net.Analysis



<http://www.netgen.com>

NetGenesis tecnologia permite a identificação de visitantes e clientes e transforma o seu comportamento em informação útil.

NetGenesis tecnologia: permite compreender o comportamento dos utilizadores nos numerosos *Web sites*.

Suporta todas as técnicas de recolha de dados e técnicas de leitura de ficheiro de registo de acessos.

Suporta centenas de site com grande volumes de visitas e com grande escalabilidade

Características da solução de E-business:

- E-metrics quantifica o desempenho de e-business;
- Identificação única de indivíduos individuais;
- Perfil dos visitantes e previsão do seu comportamento na *Web*;

Eficácia dos seus afiliados

CustomerCentric permite avaliar a eficácia dos seu afiliados capturando o identificador de um utilizador específico ligado a um *website* de um parceiro, e determina os canais mais lucrativos, e seguir os seus clientes mais lucrativos e identificar os servidores de conteúdos que eles procuram. Sendo assim o *CustomerCentric* permite:

- Efectuar relatórios sobre o número de pessoas que visitaram o site, o tempo que ficaram e qual a percentagem de utilizadores que se converteram a clientes.
- Análise dos servidores de conteúdos.
- Identificação dos clientes mais lucrativos.
- Seguir os clientes mais lucrativos e executar campanhas directamente direccionadas para esse clientes.
- Articular o ROI com as renovações de contractos e os novos contractos resultantes das campanhas de *marketing*.

Eficácia do Comércio

O *site* que tem como objectivo as transacções comerciais querem que os seus utilizadores se registem, subscrevam, façam *downloads*, comprem os seu produtos ou que façam alguma acção que lhes permita quantificar e optimizar os seus negócios.

O *CustomerCentric* permite compreender as circunstancias que levam a um tipo de comportamento desejável e avaliar a eficácia das transacções. Compreender como os visitantes se tornam em clientes.

Eficácia dos conteúdos

CustomerCentric permite analisar quais os conteúdos que são mais atractivos para os clientes. E quais os autores que são mais populares com determinado segmento de clientes. Sendo assim podemos:

- Aplicar um determinado conteúdo a um cliente individual ou a um segmento de clientes.
- Executar análises de utilizadores baseados no conteúdos.
- Reduzir os custos através do planeamento e minimizar o redesenvolvimento do *site*.

Identificação dos utilizadores

O *CustomerCentric* permite:

- Identificar os utilizadores durante a visita ao *site*.
- Construção das definições de segmentos de clientes.
- Compreender os tipo de visitas.
- Compreender os padrões de navegação no *site*.
- Comparar os padrões de comportamento com os novo utilizadores.

Perfil de visitante

Aumentar o lucro dos cliente e ROI da *Web* com os perfis e as previsões dos dados, podendo antecipar as necessidades dos visitantes. *CustomerCentric* desenvolve perfis baseados no comportamento dos visitantes e identifica os conteúdos mais importantes e as áreas chave para os visitantes.

- Aumenta a conversão dos visitantes em clientes.
- aumenta a lealdade dos clientes.

Outras Características

- A origem de visitas específicas e a eficácia de campanhas.
- O ROI produzido por parcerias ou anúncios.
- Quais as visitas feitas pelos clientes mais lucrativos
- Quais são as ferramentas ou aplicações que atingiram os objectivos de negócio.
- Aplica técnicas para a previsão de comportamentos de indivíduos
- Determina e recomenda o que vai ser aceite por um indivíduo.
- Tem uma tecnologia que lhe permite mandar recomendações para o ecrã do cesto de compras, em tempo real.
- quantifica o sucesso das iniciativas de *cross-selling*.

NetTracker Family



<http://www.sane.com/products/NetTracker/>

Análise do Visitante

- Identifica onde é que os visitantes abandonam a sua visita e porque;
- Compreensão do comportamento e das lealdades visitante/clientes e como aumentar o seu número;
- Análise do comportamento que leva a alterações do *Web site*;
- Quantificação dos visitantes: Quantos visitam site que são novos, frequente cliente;
- Ver os dados do cliente (nome, telefone, endereço, e-mail) directamente nos relatórios do NetTracker juntamente com o seu comportamento na *Web*;
- Segmentos de visitante ordenados por: comportamento, padrões de compras, lucro, e dados demográficos;
- Identificação dos segmentos de clientes mais lucrativos;

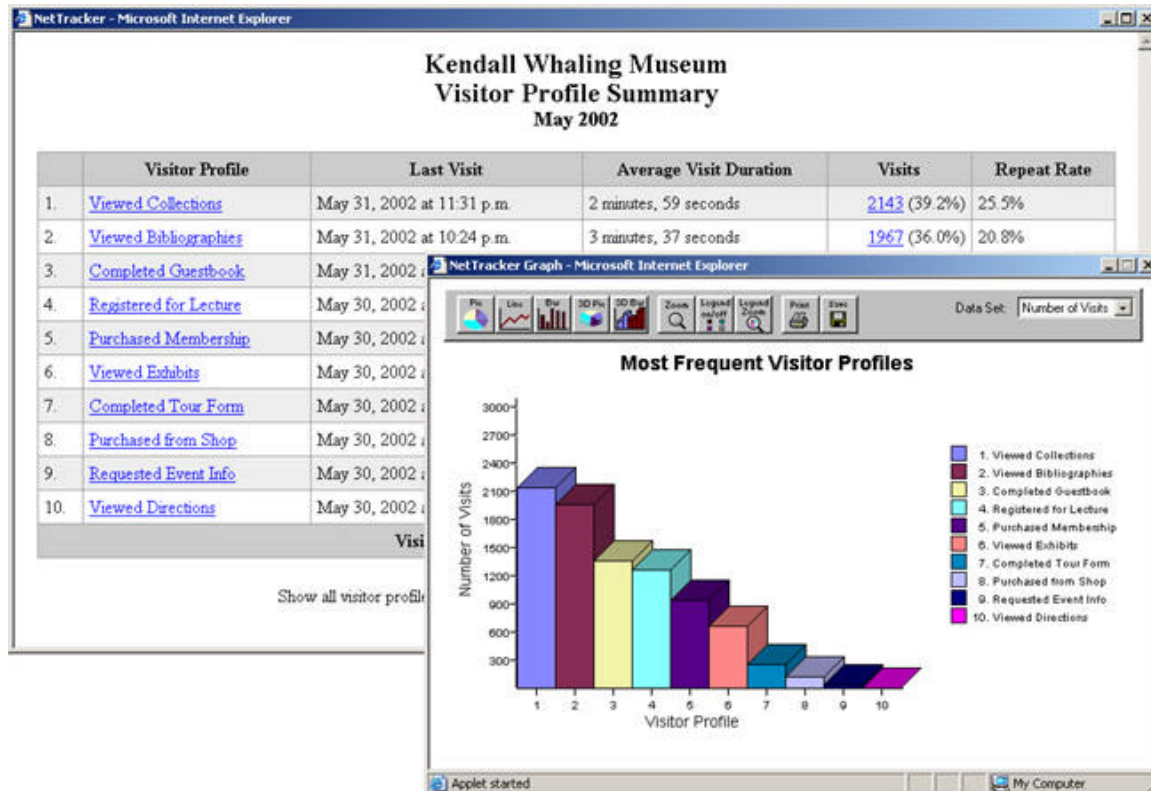


FIGURA 29: SUMÁRIO DO PERFIL DO VISITANTE DO NETTRACKER

Análise de conteúdo

- Avalia a popularidade dos produtos *Web site*.
- Identifica as necessidades dos clientes em conteúdo.
- Análise dinâmica do conteúdo do cesto de compras – quais os produtos que são frequentemente vistos online, qual o conteúdo que é rapidamente abandonado.
- compreender como os utilizadores utilizam o *Web site*.
- Eliminar conteúdos não utilizados.
- Ver os nomes dos produtos nos relatórios do NetTracker desde os produtos servidos pelos servidores quer pelos catálogos online.
- Incorporar o *feedback* do utilizador nas modificações do *Web site*.

Análise de Marketing

- Ajuda a fazer o orçamento de marketing e de alocação de decisões
- Elimina anúncios não eficazes
- Identificas os motores de pesquisa e as palavras mais procuradas no site, de forma a tomar decisões de compra e melhorar a optimização do motor de pesquisa.
- Análise referente aos clientes mais lucrativos
- Assegura que o *Web site* está a ser pelos os motores de pesquisa.
- Analisa as taxas de conversão dos visitantes durante uma visita determinados as mensagens que funcionam
- Análise as actividades trazem maior número de clientes – *banners* de publicidade, *links*, promoções por *e-mail*.

Análise de e-commerce

- Análise os padrões de tráfego para saber quais os produtos que são mais visitados
- Compreender a frequência visitas a uma site antes de comprar.
- Comparar os padrões de navegação dos visitantes que compram *online* com os que compram *off-line*.
- Análise dos lucros das compras dos catálogos *online*.

- Identifica os visitantes que vem os produtos, mas não os compram o online; cria uma lista para serem alvos de campanhas.

Análise Técnica

- Segue os *links* que estão partidos e os pedidos páginas que não foram respondidos.
- Assegura que o site é otimizado de forma a se visto pelos os *browsers* das diferentes plataformas.
- Identifica o acarreçamento lento de uma página de forma a evitar abortar o pedidos de páginas.
- segue o desempenho do servidor durante diferentes períodos de tempo.
- Diagnostico de informação sobre as actividade de cada servidor.

Análise de trafego

- Análise as visitas por data de forma a ver os resultados de uma campanha de marketing desde do dia de lançamento e durante todo o seu ciclo de vida, relacionando com a taxa de vendas.
- Segue os page views do *Web site* para alterar a quantidade de conteúdo visto durante a visita.
- Analisa as visitas por dia ou hora para determinar as tendências da utilização do site.

Análise do comportamento online dos empregados

Os relatorios do servidor/ *firewall* de proxy da NetTracker permite ver actividade na *Web* de cada empregados. Analisa o tempo total online, os sites visitados, as páginas vistas em cada site. Análise a da utilização por departamento, individual ou por grupo.

- Análise dinâmica da utilização da *Web* de cada empregado.
- Poder ver em modo gráfico, exportar ou e-mail os relatórios do servidor/firewall de proxy.
- A informação sobre a actividade para cada empregado na *Web* é detalhada pelo seguintes atributos: *Browser* , data, dia da semana, departamento, domínio, código da página, erro, servidor, palavras chave, página, plataforma, duração de sessão Page, site, página inicial, Trafego, utilizador.

Requisitos de sistema

Windows 95/98/Me/NT/2000/XP, UNIX, Mac OS X v10.0

⚡ Pentium (Pentium II/Pentium III recomendado).

⚡ 32 MB de RAM mínimo (64 MB recomendado).

⚡ 40 MB de espaço disco livre, mais os espaço para os ficheiros de registo de acessos.

Prudsys ECOMMINER



<http://www.ecomminer.com/index.html>

A família dos produtos ECOMMINER são baseados em algoritmos de *data mining* que analisam base de dados INTERSHOP, ficheiro de registo de acessos e outras bases de dados que contem informação sobre o comportamento de navegação e das compras efectuadas.

Todos os produtos ECOMMINER tem uma arquitectura cliente-servidor e são integralmente implementadas em Java para ser independentes da plataforma em que correm.

BusinessReports editon é uma solução para administradores que procuram reposta a perguntas como:

- Qual o total de vendas dos dois últimos dias?
- Quais os produtos que são comprados juntos?

AllAnalysis editon permite formular outras perguntas, sendo assim mas flexível que *BusinessReports edition*. Estas duas edições usam os mesmos algoritmos e tem a mesma estrutura.

O modulo SELECTION permite questões mais complexas como: Que produtos interessaram ao cliente X na ultima segunda-feira?

O modulo STATISTICS calcula essencialmente características estatísticas e permite facilidades gráficas para uma interpretação mais fácil.

O modulo BASKET ANALYSIS descobre regras de associação do tipo: “Se os produtos A e B são comprados juntos então o produto C também é comprado”. Todos os parâmetros escolhidos podem ser guardados em ficheiros que podem ser relacionados como macros para uma análise sequencial e automática.

O *FuturePack edition* é uma extensão do *AllAnalysis edition* adicionado-lhe avançadas funcionalidades de *data mining*. Inclui módulos para o prognóstico de futuros comportamentos de clientes bem como funcionalidades. Esta ultima funcionalidade é realizada pelo modulo TIME PREDICTOR que se baseia nas características do passado para prever o futuro. Por exemplo “Quantos produtos A serão vendidos na próxima semana?”

O modulo DIRECT MARKETING permite dirreccionar as campanhas de marketing baseado nos resultados das campanhas anteriores.

O modulo CUSTOMER PROFILER descobre os grupos de clientes que tem um comportamento de compras similares.

O modulo DATA BOOSTER é usado para incluir dados adicionais, dos clientes ou produtos de bases de dados externas para os modelos ECOMMINER e assim aumentando a precisão dos algoritmos.



FIGURE 30: CAMPOS RELACIONADOS COM O DATA MINING.

O ECOMMINER é baseado em cinco princípios básicos: flexibilidade, extensibilidade, abertura, modularização, and escalabilidade. O ECOMMINER utiliza um arquitectura aberta que é baseado no EDK (*ECOMMINER Developer Kit*).

Requerimentos de sistema

O ECOMMINER foi completamente implementado em Java e é totalmente independente da plataforma. No caso do *BusinessReports edition*, é um Java SDK (Sun Development Kit) 1.2 plug-in tem que ser instalado no *browser* do cliente.

Necessita de 64MB de RAM (128 MB são recomendados). o *FuturePack edition* requiere uma base de dados SYBASE adicional.

XAffinity(TM)



<http://www.xore.com/>

XAffinity® foi desenhado para analisar os padrões na associações e sequências.

Integração com bases de dados

- a extracção de conhecimento é feita directamente nas tabelas ou views na base de dados.
- os resultados da extracção de conhecimento é colocada directamente na base de dados

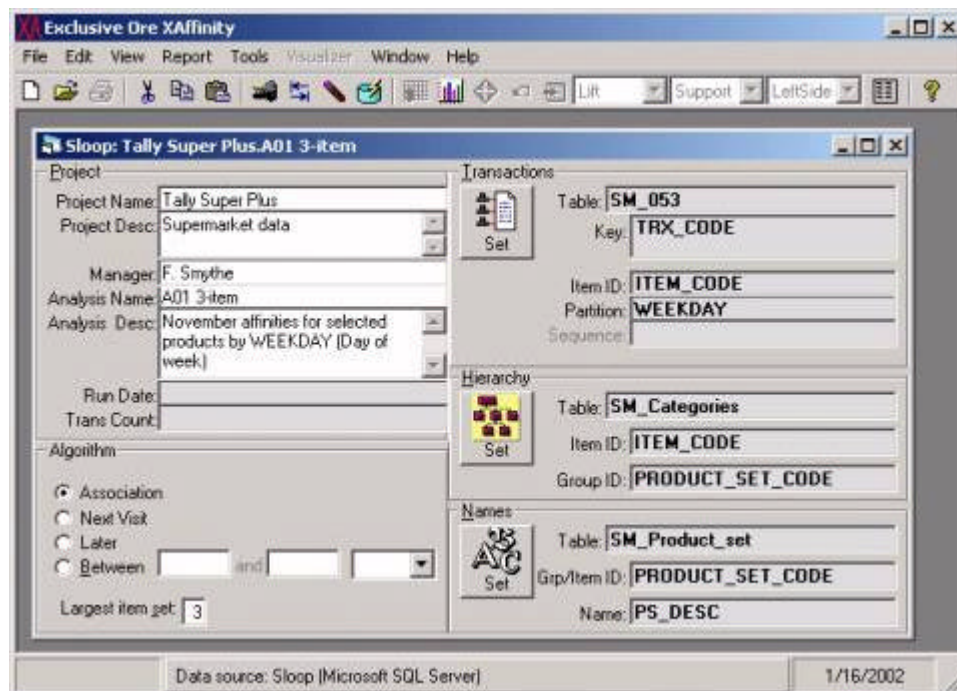


FIGURA 31: INTERGRAÇÃO DOS DADOS

Vantagens

- Podem ser utilizadas outras ferramentas, como o OLAP, para examinar os resultados.
- Muita da actividade de extracção de conhecimento decorre na base de dados.

- A segurança do DBMS é preservada porque não é necessário nenhum estrato da base de dados
- Com os resultados do modelo na base de dados, podendo aplicar os modelos a novos dados utilizando operações da base de dados (select e join) ou outras ferramentas da base de dados (OLAP).

Características do XAffinity®

- **Partição automática** – Permite automaticamente gerar regras por atributo, sendo estas guardas por: hora, dia.
- **Gerador de regras** – O gerador de regras compara as regras para ajudar a encontrar as regras com a maior (ou menor) variabilidade, exemplo: em lojas, períodos de tempo, tipo de cliente.
- **Visualização e exploração interactiva de regras** - O XAffinity® suporta a exploração de regras interactiva de um conjunto de regras. As regras podem ser vistas e exploradas em múltiplas dimensões, medidas e itens.

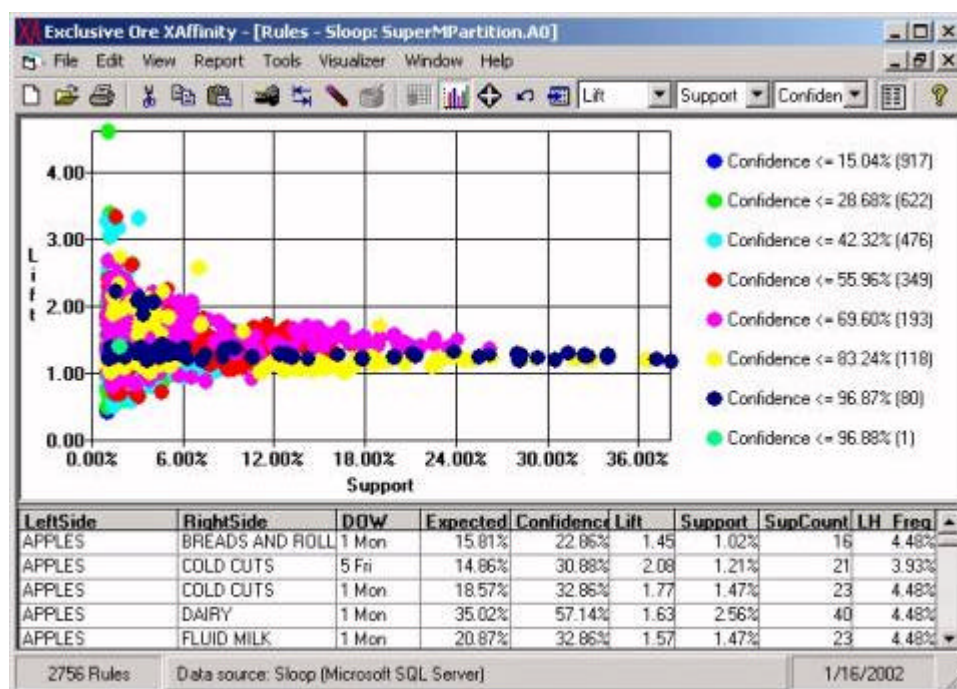
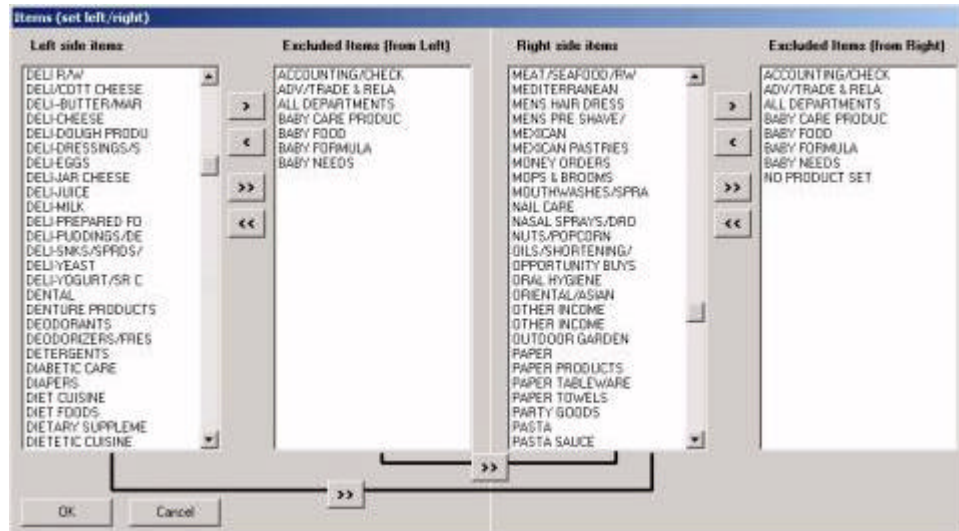


FIGURA 32: EXPLORAÇÃO DE REGRAS

- **Acesso directo aos dados** – A extracção de conhecimento é feito directamente nos dados, eliminando a necessidade de reformatar os dados.
- **Hierarquia de suporte** - A extracção de conhecimento pode ser feita nos produtos ou na página de Web, categorias ou níveis departamento.
- **Geração selectiva de regras** – As regras podem ser geradas selectivamente para itens específicos.

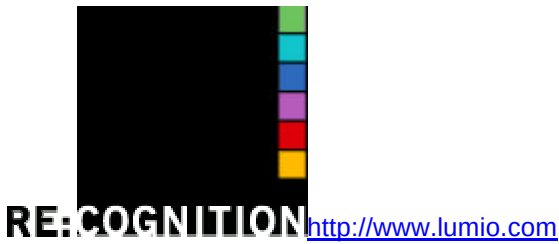


- **Múltiplos formatos de regras e texto parametrizado de regras** – As regras são produzidas em formatos especificados pelo utilizador. Os formatos da linguagem natural fazem com que as regras sejam fáceis de compreender.

As regras produzidas são dependentes do tempo e suportam:

- **Dependências de múltiplos períodos de tempo** - por exemplo mais tarde, próxima visita ou intervalos de tempo.
- **Intervalos flexíveis** – Os intervalos de tempo podem ser medidos Interval times can be measured in seconds, minutes, hours, days, months or years, or other arbitrary user-specified units.

XAffinity®, com integração com bases de dados, corre em *Windows 2000*, *Windows NT*, *Windows 98* e *Windows 95*.



Lumio Re:cognition suite foi desenhado para responder as seguintes perguntas:

- ⚡ Qual o comportamento dos clientes?
- ⚡ O que o cliente quer?
- ⚡ O que o cliente precisa?
- ⚡ Como manter os clientes?
- ⚡ Como recompensar os clientes?

Re:cognition suite inclui produtos para recolha de dados para determinação dos perfis de cliente, análise *off-line* e *real-time* dos dados.



WebHound permite seguir deixados pelos visitantes do *Web site*, permitindo assim determinar quais as páginas que atraíam mais os visitantes, e os pontos que em que os clientes saíram do site. Permite capturar todos os dados gerados pela visita ao site, conjuntamente com os dados da transacção. Sendo assim o WebHound permite:

- ⚡ Determina quais as partes do *Web site* que são mais usadas, quais as partes que os utilizadores acham mais (ou menos) interessantes, e quais as partes que estão demasiado longe para os utilizadores as encontrarem.
- ⚡ Identifica as páginas que os visitantes frequentes retornam e determina os seus pontos de interesse.
- ⚡ Determina porque os potenciais compradores abandonam as transacções.



Profit from Intelligent Information™

<http://www.torrent.com/news/torrentacquisition.html>

A tecnologia Torrent tem uma estrutura escalável de processamento paralelo para o desenvolvimento e execução de armazéns de dados, negócios inteligentes e aplicações analíticas. O processamento em paralelo aumenta o desempenho do *software*.



Analog <http://www.analog.cx/>

Analog descobre os padrões no servidor *web* procurando ser: Ultra-rápido, escalável, configurável, com relatórios em 30 línguas, podendo trabalhar em qualquer sistema operativo, e é um *Free software*

Projecto WUM (*Web Utilization Miner*)

O objectivo principal do WUM é analisar o comportamento de navegação dos utilizadores do *web site* sendo apropriado para descobrir padrões e sequências em qualquer tipo de log.

WUM é um ambiente para preparação de logs, pesquisa e visualização. Tem uma linguagem de pesquisa o MINT suporta a especificação de critérios padrões dominantes ou estatísticos.

O WUM tem as seguintes características:

- Relatórios web (HTML) e sumários compreensivos
- Pos-processamento dos resultados das pesquisas da linguagem MINT (filtros, ordenação, exportação)
- Exportação de páginas
- Execução em batch
- Relatórios para guardar os resultados das pesquisas e comentários

Índice Remissivo

A

ASP · 24, 62, 63

C

C5.0 · VI, 79, 89, 118
cache · IX, 58, 59, 60
CART · VII
classificação · VII, 13, 31, 40, 51,
52, 56, 68, 74, 75, 80, 122, 123
Classificação · VII, 37
Clustering · VII, 37, 75
comércio electrónico · 12, 25, 68
Comércio electrónico · VII
Comércio Electrónico · VII, VIII

D

Data Mining · IX, XI, 32, 33, 34, 35,
36, 38, 40, 41, 44, 45, 46, 47, 48,
49, 53, 65, 67, 68, 69, 71, 72, 73,
79, 86, 112, 116, 118, 136, 137
Datamart · VIII, 73

E

EBusiness · VIII

H

HTML · VIII, XI, 56, 58, 61, 62, 63,
70

O

ODBC · IX, 62

OLAP · IX, 33, 41, 42, 77

P

PageRank · 55, 56, 115
PMML · 64, 65
precisão · 75, 76, 77, 79, 81, 83,
100, 127, 149
Precisão · 76
Proxy · IX, 18

R

Redes neuronais · X

S

SQL · X, 16, 33, 62, 77, 78, 136

T

text mining · 50, 51
Text Mining · X, 50, 51, 52

V

Vortal · XI

W

Web Mining · XI, 53, 121

X

XML · XI, 56, 61, 64, 65, 70, 102,
118